AUTOMATED ANALYSIS OF QUANTITATIVE IMAGE BIOMARKERS FROM LOW-DOSE CHEST CT SCANS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Shuang Liu

August 2018

AUTOMATED ANALYSIS OF QUANTITATIVE IMAGE BIOMARKERS FROM LOW-DOSE CHEST CT SCANS

Shuang Liu, Ph.D.

Cornell University 2018

A quantitative imaging biomarker is a quantitatively measured characteristic derived from medical images, which serves as cost-effective and noninvasive tools for patient health assessment, including diagnosis and periodic screening of disease, therapy planning as well as longitudinal monitoring of treatment response.

This dissertation presents an automated framework for quantitative image biomarker measurement and evaluation from the low-dose chest CT (LDCT) scans that are acquired during the annual lung cancer screening. Four categories of quantitative image biomarkers are investigated, including breast density and gynecomastia quantification, bone mineral density (BMD), airway dimensions and pulmonary nodule classification. An anatomy directed approach is applied to the analysis of the breast region and to the biomarker measurements. The fully automated breast density assessment and gynecomastia measurements have been demonstrated to be consistent with the reading of radiologists. Fully automated BMD assessment is achieved by building upon the model-based segmentation and anatomical labeling of individual vertebral body. Statistically significant strong correlation with the gold standard reference can be obtained at all vertebral levels. A fully automated knowledge-based approach is applied to the segmentation and anatomical labeling of each airway bronchus, which enables the measurements of precise and reproducible airway dimensions. For the classification of pulmonary nodule malignancy, a 3D CNN is trained from scratch and

demonstrates various advantages over both the traditional machine learning approaches using hand-crafted 3D image features and the 2D CNN models. Classifier ensembles of the combinations of the 3D CNN and traditional machine learning models achieve the best performance by taking advantage of the complementary characteristics of the traditional models and the CNN models.

In conclusion, with the recent large-scale implementation of annual lung cancer screening in the US using LDCT, great potential emerges for the concurrent extraction of quantitative image biomarkers from different regions in the chest, which are covered in LDCT. This dissertation has demonstrated the feasibility of fully automated measurement and evaluation of a rich set of quantitative image biomarkers, and the opportunity to significantly enhance the impact of LDCT by offering a comprehensive health assessment to each screening participant with no additional imaging or radiation exposure.

BIOGRAPHICAL SKETCH

Shuang Liu received her B.S. degree in Electrical Engineering from Zhejiang University, China, in 2011, and her M.S. degree in Electrical Engineering from Stanford University, Stanford, California, in 2013. She became a Ph.D. student and joined the Computer Vision and Image Analysis Group (VIA) under the supervision of Professor Anthony P. Reeves.in 2013 in the School of Electrical and Computer Engineering, Cornell University, Ithaca, New York. Her research has been focusing on the areas of fully automated medical image analysis, computer-aided diagnosis, computer vision, and machine learning.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Professor Anthony P. Reeves for his insightful guidance and continuous support for my Ph.D. research and life. His exceptional knowledge, patience, and encouragement steered me in the right direction all the time for past five years. I could not have imagined having a better advisor and mentor for my graduate study.

I would also like to thank the other members of my special committee, Professor Peter Doerschuk and Professor Tsuhan Chen, for their valuable comments and motivation, which has enlightened me to broaden and improve my research from various perspectives.

A very special gratitude goes to my labmate Dr. Yiting Xie, for the stimulating discussions, passionate cooperation before deadlines, and all the funs we have had together. Without her precious friendship and support in overcoming the obstacles I have been facing in both my life and research, it would not be possible for me to complete this dissertation.

I would also like to acknowledge Dr. David Yankelevitz, Dr. Claudia Henschke, Dr. Artit Jirapatnakul, Dr. Laurie Margolies, Dr. Mary Salvatore and Dr. Emily Sonnenblick at the Icahn School of Medicine at Mount Sinai, New York. I am gratefully indebted to their valuable expertise in radiology and enlightening comments on my work.

Finally, I must express my very profound gratitude to my parents, Yanyong Liu and Yuzhi Zheng, and to my husband, Steve H. Dai, for the unfailing love and emotional support along the way throughout my years of study. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

A quantitative imaging biomarker is a quantitatively measured characteristic derived from medical images as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention [1, 2]. Similar to biospecimen-derived biomarkers, quantitative image based biomarkers can provide information on anatomically and physiologically relevant parameters [1], serving as cost-effective and noninvasive tools [3] for patient health assessment, including diagnosis and periodic screening of disease, therapy planning as well as longitudinal monitoring of treatment response [4].

An extensive variety of quantitative biomarkers can be obtained accurately and precisely from medical images with appropriate calibration, owing to the remarkable advances in medical imaging technology within recent decades [1, 2, 3]. Many imaging biomarkers are nowadays used routinely in healthcare including American College of Radiology Breast Imaging Reporting and Data System (ACR BIRADS) mammographic breast morphology that is used in breast cancer diagnosis; and bone mineral density T-score measured from dual energy X-ray absorptiometry (DXA) scans, which guides the osteoporosis diagnosis and the prescription of bisphosphonates to patients with breast cancer and bone loss induced by therapy [3].

With the rapidly increasing use of medical imaging in current clinical practice, especially imaging obtained periodically in the context of annual screening (such as screening for breast cancer, lung cancer, cervical cancer and colorectal cancer), a great deal of information needs to be interpreted comprehensively by radiologists or other medical

professionals within a short period of time. Fully automated extraction and evaluation of image biomarkers offer great opportunities to significantly enhance the impact of medical imaging [4] by eliminating unnecessary human intervention from the workflow and providing precise and reproducible quantitative information that assists the doctors in the further interpretation of medical images.

This dissertation presents a fully automated framework for quantitative image biomarker measurement and evaluation from the low-dose chest CT scans that are acquired during the annual lung cancer screening. Four categories of quantitative image biomarkers are investigated including biomarkers from the breasts, biomarkers from the bones, biomarkers from the airway and biomarkers from the lungs, as illustrated in Figure 1.1.

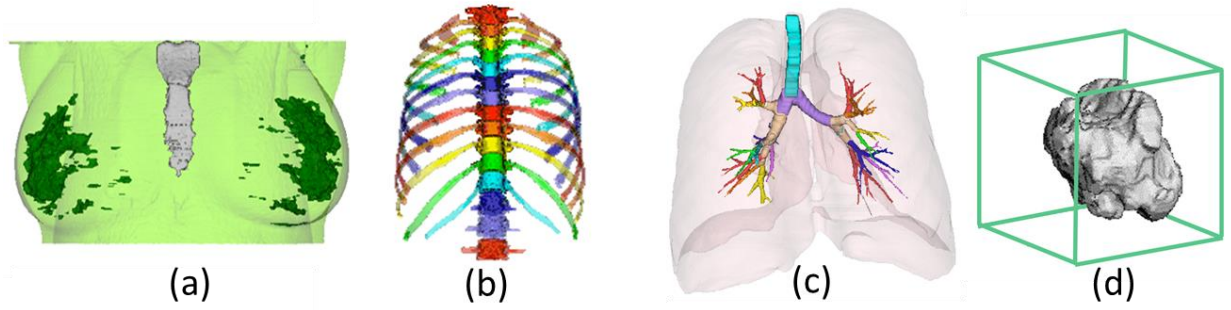

Figure 1.1. Four categories of quantitative image biomarkers are investigated including (a) biomarkers from the breasts, (b) biomarkers form the bones, (c) biomarkers from the airway and (d) biomarkers from the lungs.

1.1 Annual lung cancer screening and low-dose chest CT

Lung cancer is the leading cause of cancer death among both men and women, with more people die of lung cancer than of colon, breast, and prostate cancers combined [5]. It is estimated that each year in the US, there are 234,030 new cases of lung cancer and 54,050 lung cancer-associated deaths [5]. Since the symptoms of lung cancer typically do not appear

until the disease is already at an advanced stage, the diagnosis of lung cancer is usually delayed, and therefore more than 50% people die within one year of being diagnosed [6].

The survival for lung cancer is strongly related to the stage at which it is first diagnosed. It is demonstrated that the current 5-year lung-cancer–specific survival rate is 17.7%, while the 10-year lung-cancer–specific survival rate is 80% for people that are diagnosed with lung cancer during the annual screening [7], which are usually in curable early stage.

Annual lung cancer screening with low-dose chest CT (LDCT) has been used for early detection of lung cancer and has been demonstrated to reduce mortality from lung cancer [8] and has been covered by medical insurances since 2015 [9]. It is estimated that 8.6 million Americans meet the annual lung cancer screening criteria (current and former smokers who were aged 55 years to 74 years) and the use of LDCT is increasing rapidly [10]. As the LDCT in general covers the regions of lungs, airway, breasts, mediastinum and several bone structures (including clavicles, sternum, ribs and vertebrae), and the lung cancer screening population are also at high risk of having a variety of other diseases (such as breast cancer, osteoporosis and chronic obstructive pulmonary disease), LDCTs acquired from annual lung cancer screening potentially serve as a rich and valuable source for the automated measurement and analysis of an extensive variety of quantitative image biomarkers, as illustrated in Figure 1.1.

Lung cancer screening with LDCT is generally accomplished at an overall average effective dose below 1.0 mSv for an average-size participant as compared with an average effective dose of 7 mSv for a typical standard-dose chest CT examination used in the daily

clinical practice [11]. The image noise and absorbed dose are intrinsically linked because they both are fundamentally related to the X-ray fluence used during image acquisition [12]. Therefore, LDCT scans in general have much higher level of image noise as compared with standard dose CT, which makes fully automated analysis of LDCT scans more challenging. Therefore, the computer algorithms designed for the standard-dose CT usually do not generalize well to LDCT scans.

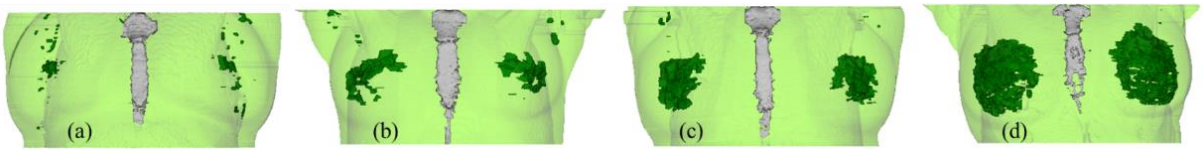

Figure 1.2 (a-d) Four cases in 3D coronal view in the order of increasing breast density where the segmented breast (light green), fibroglandular tissue (dark green) and sternum (grey) are shown.

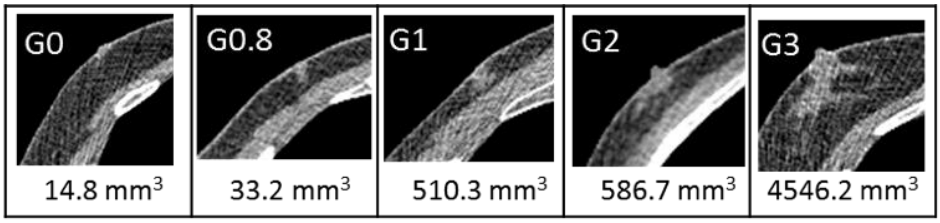

Figure 1.3 Examples of male breasts of different gynecomastia reference grades. For each figure, the reference grade is shown on the top and the automated measurement is shown on the bottom.

1.2 Automated measurement of quantitative image biomarkers from low-dose chest CT

Two quantitative image biomarkers, breast density measurement for women as illustrated in Figure 1.2, and gynecomastia quantification for men as illustrated in Figure 1.3, were investigated from the breast region. Breast density is an independent risk factor for breast cancer, which is the most common cancer (excluding skin cancers) diagnosed among US women and also the second leading cause of cancer death among US women [13]. Dense breasts have up to 6 times greater risk of breast cancer than the less dense breasts, hence

leading to legislation mandating that women be informed of their breast density on mammogram reports [14]. Breast density is usually assessed using Breast Imaging Reporting and Data System (BI-RADS) from mammography, which is currently most commonly used modality for breast imaging. The purpose of this study is to take full advantage of LDCT scans acquired in the annual lung cancer screening by providing critical measurements about breast heath without additional radiation exposure, rather than to use LDCT to replace mammography.

The second investigated quantitative image biomarker from the breast region is the quantification of gynecomastia, which is characterized by the benign enlargement of male breasts due to the growth of glandular tissue. It is a common and sometimes distressing condition found during physical exam in over half of normal adult men over the age of 44 [15, 16]. Although the majority of gynecomastia is physiologic (i.e. associated with puberty or aging) or idiopathic (i.e., cause is unknown) [15], its occurrence may also associate with an extensive variety of underlying systemic disease or drug toxicity [15, 17]. Reliable gynecomastia quantification can assist the early detection as well as the treatment of both gynecomastia and the underlying medical problems, if any, that cause gynecomastia.

The assessment of bone mineral density (BMD) plays an essential role in the diagnosis and follow-up therapy monitoring of osteoporosis, which is the most common metabolic bone disease and is estimated to affect 12.3 million US population aged 50 years or older in 2020 [18]. Osteoporosis is characterized by low bone density and micro-architectural deterioration of bone tissue [19], with related complications, such as osteoporotic fractures, becoming a significant cause of increased morbidity and mortality [20], and thereby creating tremendous social and economic burdens. This study investigated the BMD measurements and

demonstrates the potential of opportunistic osteoporosis screening with concurrent lung cancer screening using LDCT.

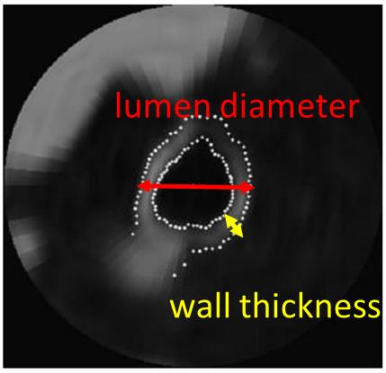

Figure 1.5 Illustration of airway dimension derived quantitative image biomarkers, including lumen diameter and wall thickness.

Airway dimension derived quantitative image biomarkers, including lumen diameter and wall thickness as illustrated in Figure 1.5, have been demonstrated repeatedly to correlate well with the severity of airflow obstruction and peripheral airway inflammation in patients with chronic obstructive pulmonary disease (COPD), which is a heterogeneous disease associated with varying degrees of emphysema and small airways disease [21] and is expected to be the 3rd leading cause of death by 2020 [22]. As a consequence, the precise and reproducible measurements of airway dimension derived biomarkers may facilitate more accurate COPD diagnosis, treatment planning, as well as the evaluation of therapy response [21].

Pulmonary nodules are approximately spherical regions (with diameter $\leqslant$ 30 mm) of the lungs, which usually appear on CT images as solid tissue having a much higher image intensity than the surrounding lung parenchyma [23], as illustrated in Figure 1.6. Sub-solid nodules, also known as ground-glass opacities, which grows along the air-containing

structures of the lung and has a nonsolid appearance on CT images [23], are much less frequent thus are not considered in this study. The detection and malignancy classification of the pulmonary nodules are essential during the annual lung cancer screening, as the nodules can be benign (noncancerous) that often requires no treatment, or malignant (cancerous) that requires medical treatment as soon as possible. The analysis of pulmonary nodule biomarker in this study focuses on the prediction of the malignancy status from the initial CT scan finding to improve the costly follow-up procedures.

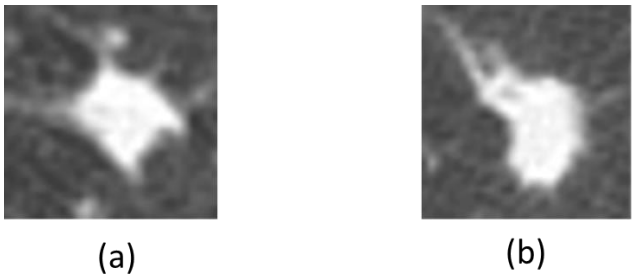

Figure 1.6 Examples of (a) malignant nodule and (b) benign nodule in axial view from LDCT scans.

1.3 Overview

A fully automated framework is presented in this dissertation for the measurement and evaluation of quantitative image biomarkers from LDCT scans acquired during the annual lung cancer screening. Quantitative image biomarkers from the following four categories of were investigated:

(i). Breast region analysis and quantitative image biomarker measurement in Chapter 2. In order the determine the region of interest for the subsequent biomarker measurements, fully automated anatomy directed algorithm is first developed for the segmentation of the whole breast region and the fibroglandular tissue. Then a machine learning based approach is used to for the localization of nipples. Finally, the two quantitative image biomarkers, including

breast density for women and gynecomastia quantification for men, are measured and validated by comparing to the reference established by the radiologists.

(ii). Bone structure analysis and quantitative image biomarker measurement in Chapter 3. Individual bone structures, including clavicles, sternum, ribs and vertebrae, are first segmented sequentially and anatomically labeled by a fully automated model-based algorithm. Then building upon each of the segmented and labeled vertebra, the BMD is measured and validated by using the BMD measured form the DXA scans the reference.

(iii). Airway anatomical labeling and quantitative image biomarker measurement in Chapter 4. Each individual airway bronchus is first segmented and labeled with its anatomical name using a knowledge-based approach. Then two airway dimension derived quantitative image biomarkers, the lumen diameter and wall thickness, are measured and validated based on the reproducibility of the measurements.

(iv). Pulmonary nodule malignancy classification in Chapter 5. A 3D convolutional neural network (CNN) trained from scratch is employed for the classification of pulmonary nodule malignancy. Classifier ensembles of different combinations of the 3D CNN and traditional machine learning models based on handcrafted 3D image features are also explored.

CHAPTER 2

FULLY AUTOMATED BREAST ANALYSIS AND QUANTITATIVE BIOMARKER MEASUREMENTS

Fully automated breast health analysis from LDCT scans acquired during the annual lung cancer screening is considered in this chapter. For female subjects, breast density is quantified and can potentially be used for breast cancer risk estimation. For male subjects, the gynecomastia quantification is measured and can potentially be used for the gynecomastia diagnosis and treatment planning. The presented framework is the first published work on the fully automated breast density measurement and gynecomastia quantification from LDCT, demonstrating the feasibility of concurrent breast healthy analysis for both women and men and annual lung cancer screening without requiring any additional patient time or radiation exposure.

Breast cancer is the most common cancer (excluding skin cancers) diagnosed among US women, with approximately 232,340 new cases [13] accounting for nearly 1 in 3 cancers in the US each year. It is also the second leading cause of cancer death among US women which results in 39,620 deaths in the US each year [13]. Breast density has been shown to be an independent risk factor for breast cancer, hence leading to legislation mandating that women be informed of their breast density on mammogram reports in many states [14].

Mammography is recommended for breast cancer screening by many organizations with most US mammograms being 2D full field digital mammograms (FFDM). Digital breast tomosynthesis and 3D imaging modalities, including CT [24, 25, 26], ultrasound [27] and MRI [28] are of increasing interest. 3D modalities are free of the inherent superimposition

effects of mammography, which is a 2D projection of 3D breast volumes, thereby being helpful especially for women with dense breasts, for which mammographic screening may have relatively low sensitivity [24, 25, 29]. Several recent studies [24, 25, 26, 30, 27] have demonstrated that breast density readings on 3D modalities are consistent with readings on 2D mammogram with comparable inter-observer agreement.

Breast density is usually assessed using Breast Imaging Reporting and Data System (BI-RADS) that classifies breast composition into 4 categories [28]. Percentage quartiles are no long employed for the division of the density groups in the latest BI-RADS atlas to better reflect the masking effect of dense fibroglandular tissue and estimate the density volume [28]. There are currently no published studies to validate a computer aided breast density assessment framework for LDCT by comparing with the subjective grading of radiologists following the latest BI-RADS guidelines.

Several recent publications have demonstrated that breast density measured from LDCT correlates well with density measurements from mammogram and MRI [24, 25]. Consequently, LDCT can potentially serve as a valuable resource providing useful information with respect to breast cancer risk evaluation for many women who have undergone LDCT but not recent mammograms.

Gynecomastia is characterized by benign enlargement of male breasts, as shown in Figure 2.1, due to the growth of glandular tissue [31], which is a common and sometimes distressing condition found during physical exam in over half of normal adult men over the age of 44 [15, 16]. Most cases of gynecomastia are due to an imbalance in estrogen and androgen action, with estrogen-induced stimulation predominating [15].

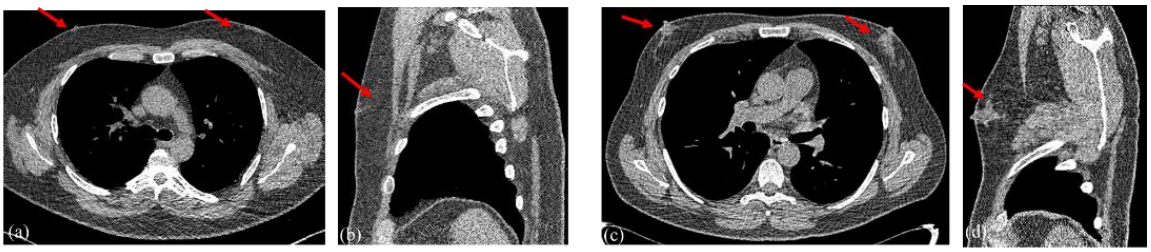

Figure 2.1. Normal male breasts in (a) axial view and (b) sagittal view and male breasts with gynecomastia in (c) axial view and (d) sagittal view. The subareolar regions are indicated by arrows.

Although the majority of gynecomastia is physiologic (i.e. associated with puberty or aging) or idiopathic (i.e., cause is unknown) [15], its occurrence may also associate with an extensive variety of underlying systemic disease or drug toxicity [15, 31]. Physiologic gynecomastia is often an incidental finding and may be painful or cosmetically disabling in some men, who may hence benefit from treatment [31]. Drug-induced gynecomastia may be treated with discontinuation of an offending drug, if it is identified during the proliferative phase (i.e., gynecomastia has been present within a year) [32]. On the other hand, if the gynecomastia is of long duration, the breast enlargement is unlikely to regress substantially, due to presence of fibrotic tissue [32].

Gynecomastia must be distinguished from pseudo-gynecomastia [33, 31], which is characterized by increased subareolar fat without enlargement of the breast glandular component, and requires no further investigation [33]. Although rare in men, gynecomastia also needs to be distinguished from breast carcinoma in the differential diagnosis [15]. Physical examination without imaging is usually sufficient for the diagnosis of gynecomastia [32, 34].

The radiographic appearance of gynecomastia on CT scans has been investigated in several studies [34, 35, 36], which are all based on subjective reading by radiologists.

Although the appearance of gynecomastia has been well described on mammography, ultrasound and digital breast tomosynthesis [34, 29], there is no consensus on the standard measure for the diagnosis and quantification of gynecomastia [34]. In the study [17] by Cooper et al, gynecomastia was quantified using a four-category grading scale based on the volume fraction of dense tissue in the subareolar region: gynecomastia was considered mild when dense tissue occupied less than 25% of the volume of the subareolar soft tissue; moderate when it occupied 25-50% of the subareolar volume; and extensive when it occupied more than 50% the subareolar volume. Other studies [34, 36] have employed the maximal diameter measured at the axial level of the nipples as the quantity measure of gynecomastia. A subareolar glandular tissue diameter of 20 mm and above on CT scans was considered as gynecomastia by Klang et al in [36], which is consistent with the diagnosis criteria used in the physical examination by Nuttall et al in [16].

As the LDCT in general also covers the regions of breasts and the high prevalence of gynecomastia among the lung cancer screening population (current and former smokers who were aged 55 years to 74 years), gynecomastia is believed to be a frequent incidental finding [34] on LDCT scans acquired during annual lung cancer screening. Therefore, LDCTs can potentially serve as a valuable resource for the early detection and longitudinal monitoring of gynecomastia, which may aid the treatment of both gynecomastia and the underlying medical problems, if any, that cause gynecomastia.

The locations of the nipples are important anatomical landmarks as they are the only consistent spatial references of human breast [37, 38]. Radiologists often use nipple location as the reference point to aid the interpretation of breast imaging [34, 35, 34], including breast density assessment and gynecomastia quantification. Computer aided analysis of breast

imaging frequently requires automated nipple localization as the first step. The applications [37, 39, 40] of fully automated nipple localization include but are not limited to: (1) registration of longitudinal breast imaging of the same subject for the monitoring of changes occurred in breast; (2) registration of left and right breast imaging for the identification of bilateral asymmetry which can be a sign of breast cancer [41]; (3) registration of multi-view 2D projections or different imaging modalities to aid the diagnosis of breast abnormalities [42, 43], such as masses and microcalcifications; (4) construction of an anatomical frame of reference facilitating the segmentation of anatomical structures, such as pectoral muscles, fat tissue and glandular tissue, and breast density quantification [43, 44].

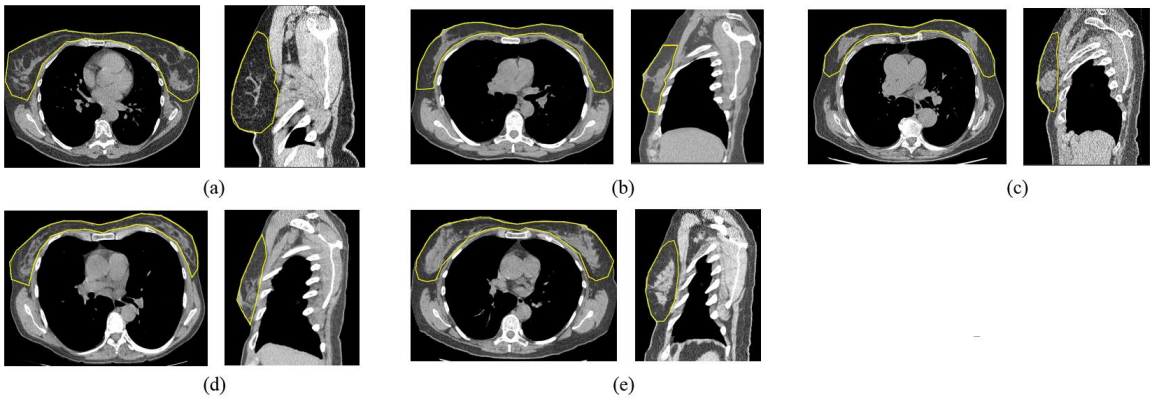

Figure 2.2. Five examples of breast region with ground truth annotated (indicated by yellow contours) by a radiologist. For each case, both an axial slice and a sagittal slice of the left breast are shown. It can be seen that there is a large range of individual variations of breasts.

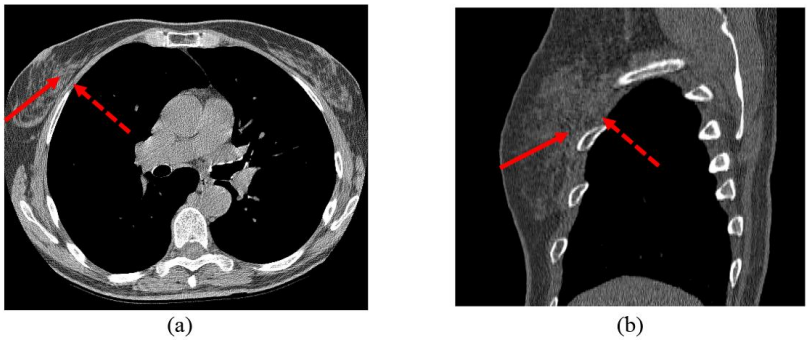

Figure 2.3. CT (a) axial slice and (b) sagittal slice. Solid arrows mark glandular tissues in the breast and dashed arrows mark the muscles in the pectoral regions.

A fully automated framework for the breast health analysis from LDCT has been developed. The whole breast region is first segmented using an anatomy-orientated approach, in which the vertical, anterior and lateral extents of the breast are determined based on the spatial constraints defined by other human tissues and organs, and the posterior extent is resolved by the propagation of the pectoral muscle fronts as the separation between the breast region and the underlying muscles. The subareolar region is then localized using a machine learning based nipple detection algorithm. Finally, building upon the region of interest defined by the segmented breast and nipple location, two quantitative image biomarker measurements, the breast density for women and gynecomastia quantification for men, are accomplished.

Automated breast region analysis on CT images have been presented in several previous studies [24, 25, 44, 43, 45, 46, 47], which mainly focused on the breast density assessment for female breasts. The segmentation of fibroglandular tissue in the breast region usually serves as the first step towards automated breast analysis. Semi-automated template-based segmentation methods have been presented by Chen et al [24]and by Moon et al [25] for LDCT, where manual annotations are needed for the generation of a template slice for each scan. Fully automated anatomy-orientated knowledge-based segmentation methods have been presented in our own work [45, 46] for LDCT and by Zhou et al [44] for regular dose torso CT. A machine-learning based approach has also been employed for the segmentation of fibroglandular tissue by Zhou et al [47] for regular dose torso CT. Zhou et al [43] recently applied a deep convolutional neural network for the fully automated direct classification of female breasts into one of the four density categories without the need of fibroglandular tissue segmentation on regular dose torso CT. No previous studies have been found to date on the

fully automated breast region analysis of men breasts or gynecomastia quantification on CT scans.

A number of studies have been conducted on the automated nipple localization [37, 38, 39, 40, 41, 42, 44, 48], while most of the algorithms were proposed for mammography [37, 38, 39, 40, 41, 42] that is a 2D imaging modality. A semi-automated nipple localization approach was presented by Gwo et al [48] for breast MRI, using breast imaging coils for females in a prone position. For CT, chest images are usually acquired in a supine position. Zhou et al [44]proposed a method for nipple localization from standard-dose torso CT as part of the breast segmentation algorithm based on the analysis of skin surface curvature and the tissue composition near the skin surface. However, the performance of the nipple localization algorithm was not reported.

2.1 Breast segmentation and nipple localization

There are two main challenges to the automated segmentation of the whole breast in LDCT. First, the algorithm must accommodate a significant range of individual variations in terms of size, shape, location and tissue compositions of the breasts as illustrated in Figure 2.2. LDCT scans are usually taken in the supine positions, while other modalities such as mammogram, dedicated breast CT, and breast MRI acquire images either in the prone position or with breast compression to constrain the location and shape of the breasts. As a result, there is a much greater range of variations in terms of shape and location of the breast in LDCT images than in images of other modalities. Second, glandular tissues located in the posterior breast regions may be difficult to be distinguished from the surrounding muscles in the pectoral regions as illustrated by arrows in Figure 2.3, because they have similar CT intensity distributions and can be in contact with each other, thereby lacking in well-defined

boundaries to exclude the muscles from the breast regions. Moreover, the glandular tissues usually have irregular shapes, which complicates the task of separating glandular tissues and muscles.

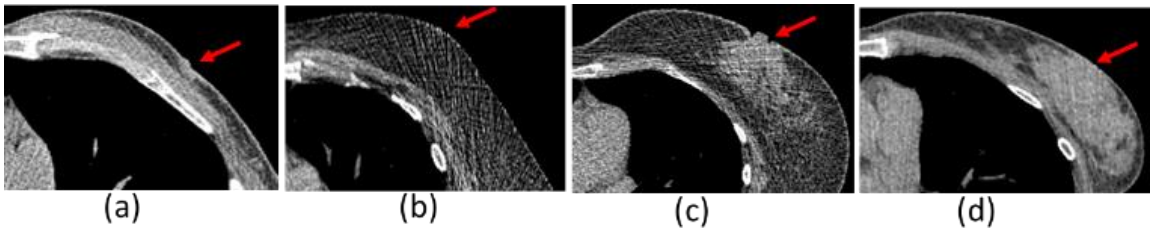

Figure 2.4 Variations of nipples (marked by arrows) across individuals.

Fully automated nipple localization from LDCT is challenging for two primary reasons. First, LDCT scans usually exhibit a much higher level of image noise compared to standard-dose CT [49], as a much lower overall radiation dose is adopted. Second, there is a significant range of variations in the appearance, size, shape, and location of nipples across individuals as illustrated in Figure 2.4. While one might expect that the nipple would be identified by a "bump" on the surface of the skin pointing out from the body as shown in Figure 2.4 (c), a number of alternative presentations are observed in CT images. For example, the nipple region may be: smooth as shown in Figure 2.4 (d), inverted as shown in Figure 2.4 (a), or lacking any glandular tissue and being smooth as shown in Figure 2.4 (b).

The main contribution of the presented breast segmentation approach is the novel algorithm, muscle front propagation, for the separation of pectoral muscles and fibroglandular tissue. It is designed to model and track the muscle front that is more well-defined in terms of shape and location compared to the fibroglandular tissues that often have large individual variations of in terms of position, volume and shape as indicated in Figure 2.7. It takes

advantage of the fact that the muscle front is generally of convex shape and smooth both in the vertical direction and the on the axial plane. The main contribution of the nipple localization algorithm is the application of machine learning based method and the extraction of effective LDCT image features of nipples. Moreover, it is the first published work on the fully automated nipple localization from LDCT scans.

2.1.1 Segmentation of whole breast and fibroglandular tissue

The whole breast region is modeled in this paper as a two-component region, consisting of fat tissue and fibroglandular tissue, which is located outside the thoracic cavity and anterior to the chest muscle as indicated in Figure 2.5. The whole breast region is first segmented using a novel anatomy-orientated approach based on the propagation of muscle fronts to resolve the challenge of separating the fibroglandular tissue from the underlying muscle. The fibroglandular tissue and fat tissue are then identified from the segmented whole breast and the percentage breast density is calculated based on their volume ratio.

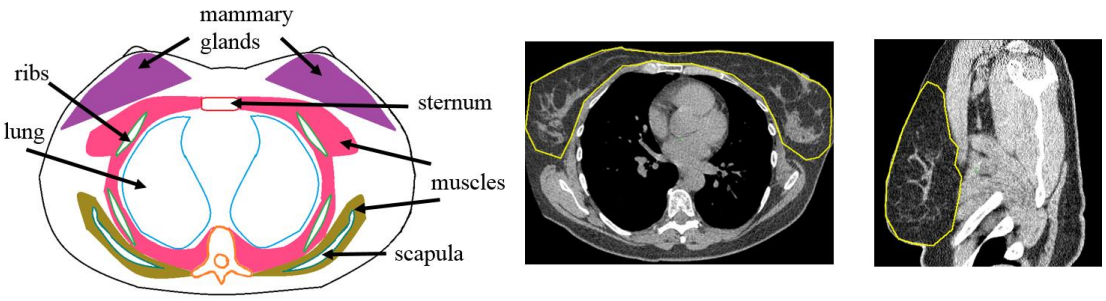

Figure 2.5 (a) Human tissues and organs in the chest region shown in an axial view. (b) Whole breast region annotated (in yellow) by a radiologist in an axial CT slice and (c) in a sagittal CT slice.

The segmentation of the whole breast requires the determination of its extents in the vertical, anterior, lateral and posterior directions. Several adjacent tissues and organs as

illustrated in Figure 2.6 (a), including sternum [50], ribs [51], lungs [23], skin, fat and soft tissues [49], are segmented by algorithms developed in our previous studies and employed as prior dependence. The presented whole breast segmentation algorithm consists of the following three main steps:

(1). Define the vertical extents of the breast as the range between the superior end of the sternum and the inferior end of the lung as shown in Figure 2.6 (b).

(2). Define the anterior and lateral extents of the breast based on the skin surface as shown in Figure 2.6 (c).

(3). Define the posterior extents of the breast based on the muscle front shown in Figure 2.6 (c) and obtained by initialization at the superior axial level and then propagation in the inferior direction, which is explained in detail.

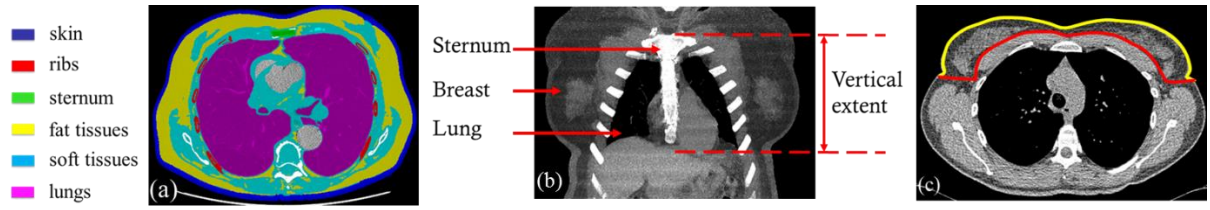

Figure 2.6 (a) Human tissues and organs used as prior dependence are shown in axial view. (b) The vertical extents of the breast are determined based on sternum and lung, shown in coronal view. (c) The anterior and lateral extents (in yellow) of the breast are determined based on the skin surface. The posterior extents of the breast (in red) are determined based on the muscle front.

The third step, muscle front propagation, is a novel algorithm proposed in this paper to serve as a solution to the well-known challenge [46, 47] of separating muscle from fibroglandular tissue as demonstrated in Figure 2.7. A Cartesian coordinate system, where the x, y and z axes are defined along mediolateral, anteroposterior, and craniocaudal directions respectively, is established as shown in Figure 2.7 to aid the following description.

The muscle front is defined in this paper as the interface of the muscles and other adjacent tissues, which are usually fat tissues or fibroglandular tissues. Since at the superior level of the breast region (i.e., z = 0, the axial level at superior end of the sternum) there are no fibroglandular tissue, the muscle front can be easily initialized based on the boundary between fat tissue and non-fat soft tissues as shown in Figure 2.8 (b) (For the remainder of the chapter, we will use the term soft tissues to refer to non-fat soft tissues.). The smoothness of the muscle front allows the front propagation along the inferior direction through the whole vertical range of the breast as shown in Figure 2.8. At the location of where the muscle front is not well defined due to the existence of adjacent fibroglandular tissues as in Figure 2.7, the front can be resolved according to the adjacent front locations both axially and vertically.

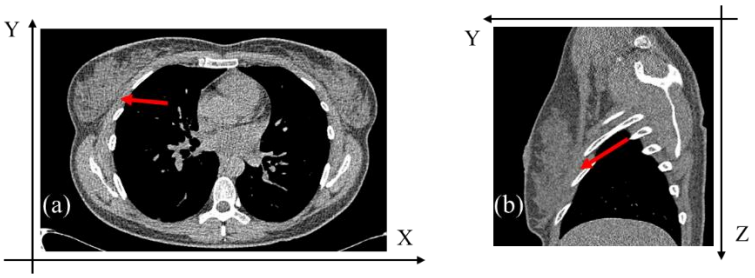

Figure 2.7. The fibroglandular tissue and the underlying muscle can be difficult to separate as indicated by arrows in (a) axial view and (b) sagittal view.

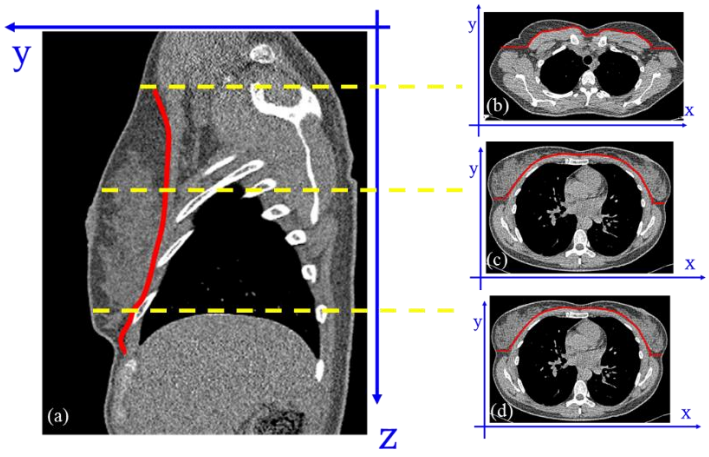

Figure 2.8. Muscle front propagation along inferior direction through the whole vertical range of the breast shown in (a) sagittal view and (b-d) axial view.

The flow chart of the muscle front propagation algorithm is shown in Figure 2.9. The remainder of the algorithm description focuses on the right hemithorax using the sternum as the medial separation; the algorithm for the left side is similar due to symmetry.

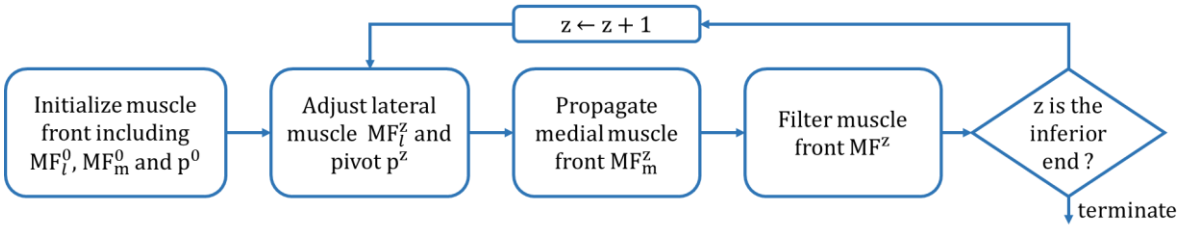

Figure 2.9. Flow chart of the muscle front propagation algorithm.

The muscle front is modeled as a surface f(x, z) that is parametrized by the x and z coordinates. It is initialized at the superior level z = 0 of the breast and propagated towards the inferior direction as z increases. As shown in Figure 2.10 (a), at each axial level z, the muscle front f(:, z) consists of two components: the medial muscle front (MF_m^z) and the lateral muscle front (MF_l^z), where a pivot $p^z = (p_x^z, p_y^z)$ is located at the border between them.

The lateral muscle front MF_l^z separates the breast region from the posterior body and eliminates the need of front propagation of muscles in the back, such as muscles attached to scapula. The lateral muscle front serves the same purpose as the medial muscle front in excluding muscles from the breast region and constituting a posterior extent of the breast region.

The medial muscle front MF_m^z is defined as the anterior interface of the muscles anterior to MF_l^z and outside the chest cavity (CC), where CC is approximated by the 3D convex hull of sternum, ribs and lungs as illustrated in Figure 2.10 (b-c). The pivot p^z is

defined as the most lateral point (with smallest x coordinate for the right size of the body) on MF_m^z.

The front candidates (FC) as illustrated in Figure 2.10 are potential locations to which the medial muscle front can propagate, and defined as the boundary pixels of the union of soft tissues and chest cavity CC. The boundary of the CC is also considered as FC in order to deal with the location on the chest wall with no significant soft tissue attached as indicated in Figure 2.10 (c).

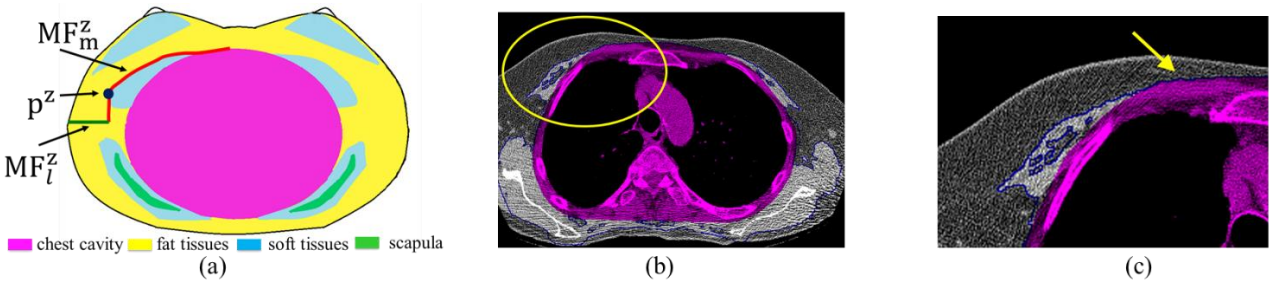

Figure 2.10. (a) Muscle front f(x,z) shown in an axial slice z. (b-c) The front candidates (FC, in blue) and the chest cavity (CC, in magenta) in (b) axial view and in (c) magnified view of the circled region in (b). The arrow points at the location where the FC lies on the boundary of CC due to the lack of attached soft tissue.

At the superior breast extent z = 0, the lateral muscle front MF_l^0 is initialized by determining the y level y_l^0 that minimizes the following score function (2.1) as shown in Figure 2.11, which aims to located a y level with minimal amount of soft tissues S(y) to avoid cutting through any fibroglandular tissue. Greater lateral body width W(y) is also encouraged based on empirical observations.

$$y_l^0 = \underset{y_p^0 < y < y_a^0}{\operatorname{argmin}} \; \alpha\, S(y) + \beta\, W(y) \tag{2.1}$$

$$S(y) = \sum_{x < x_c^0(y),\ |y'-y| < \delta_s} 1_{ST}(x, y') \tag{2.2}$$

$$W(y) = x_c^0(y) - x_l^0(y) \tag{2.3}$$

$$\alpha > 0, \ \beta < 0 \tag{2.4}$$

where α and β are weight constants; y_a^0 and y_p^0 are anterior and posterior search limits at z = 0 respectively; $1_{ST}(x, y)$ is the indicator function for soft tissues; δ_s is the constant window size for counting soft tissue pixels; $x_c^0(y)$ and x_l^0 (y) are the most lateral x coordinates of the chest cavity and the breast region at z= 0 and y.

The medial muscle front MF_m^0 can be easily identified, since at z = 0 there is no fibroglandular tissue and any segmented soft tissues are thus muscle. As indicated in Figure 2.11, MF_m^0 can thereby be determined by searching for the most anterior FC pixels at each x coordinate. Note that only regions anterior MF_l^0 needs to be considered. The pivot p^0 can also be easily determined by locating the most lateral point on MF_m^0 as shown in Figure 2.11.

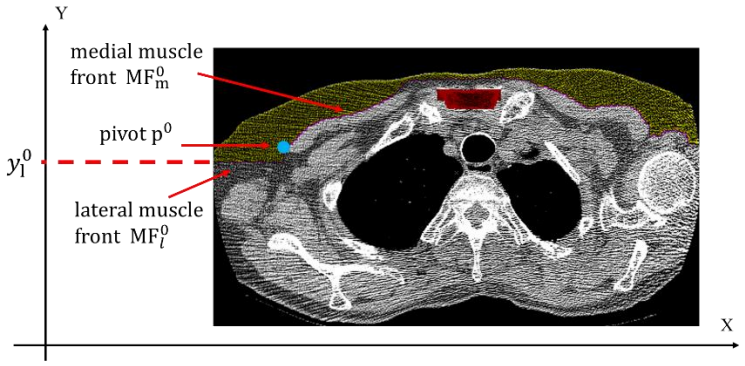

Figure 2.11. Initialization of muscle front (in red) at the superior extent (z = 0) of the breast.

As the front propagates in the inferior direction to the axial level z, the lateral muscle front MF_l^z in general needs to be adjusted from previous locations MF_l^{z-1}. A similar score function (2.5-2.6) as that defined in (2.1-2.4) is used to determine y level y_l^z with an additional term to penalize unnecessary movement from y_l^{z-1} that encourages a smooth breast boundary.

$$y_l^z = \underset{y_p^z < y < y_a^z}{\operatorname{argmin}} \ \alpha S(y) + \beta W(y) + \gamma |y_l^z - y_l^{z-1}| \tag{2.5}$$

$$\alpha, \gamma > 0, \ \beta < 0 \tag{2.6}$$

The determination of medial muscle front MF_m^z at an inferior axial level can be difficult due to the existence of fibroglandular tissues. Fortunately, the smoothness and continuity of the muscle surface suggests that MF_m^{z-1} on the previous axial level can be used as a reliable starting point for the search of the new front location MF_m^z and this search can be constrained within a limited search window. The algorithm for propagating MF_m^z at each location (x, z) is summarized in equation (2.7-2.8) below, where the front on the previous axial level f(x, z-1) serves as the starting point, and the determination of f(x, z) is in fact an optimization of the y coordinate:

$$f(x, z) = \underset{y < f(x,\ z-1) + \delta_f,\ 1_{FC}(x, y') = 1 \text{ for } \forall\, y_n < y' < y}{\operatorname{argmax}} y \tag{2.7}$$

where,

$$y_n = \underset{|y - f(x,\ z-1)| < \delta_f,\ 1_{FC}(x,\ y) = 1}{\operatorname{argmin}} |y - f(x, z-1)| \tag{2.8}$$

$1_{FC}(x, y)$ is the indicator function for front candidates (FC); δ_f is the search window along y axis, and y_n is the FC that is nearest to the start point f(x, z-1).

Note that it is possible that at some x coordinate, the muscle front f(x, z) cannot be resolved due to the lack of FC in the search window, generally resulting from closely attached fibroglandular tissue. These front locations will eventually be determined based on filtering among neighboring front locations later.

Three types of filtering are applied sequentially to the muscle front to ensure a smooth and closed posterior breast boundary:

First, the medial muscle front MF_m^z is smoothed based on the propagation speed that is measured as the location difference f(x,z) – f(x, z-1) between axial level z and z -1. An outlier is identified and marked as unresolved if its speed is significantly different from the average speed in the axial neighborhood.

Second, the whole muscle front MF^z is smoothed based on continuity that is measured as the location difference f(x, z) – f(x -1, z) in the axial neighborhood on the current axial level z. An outlier is identified and marked as unresolved if its location difference is significantly different from that in the axial neighborhood.

Finally, the y coordinates of the unresolved muscle fronts are determined by linear interpolation of the nearby resolved front locations.

2.1.2 Nipple localization

The breast is modeled by a three-compartment region for the purpose of nipple localization, consisting of skin, fat tissue and fibroglandular tissue. Three types of anatomy-oriented image features (6 feature values in total) including spatial location, depth of density and degree of protuberance are extracted for each pixel at the skin surface. Spatial location is employed due to the fact that nipple is located inferiorly to the jugular notch at the anterior front of the skin surface. Left and right nipples are approximately symmetric to the jugular notch. Depth of density is used because of the converging characteristics of breast density (non-fat tissue) at the nipple [39]. Degree of protuberance is considered based on the observation that a nipple is often characterized by a "bump" pointing out from the body or an "indentation" towards the body (for an inverted nipple as in Figure 2.4(a)).

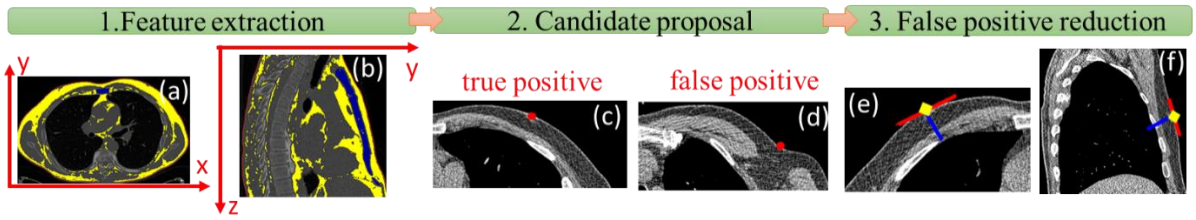

Figure 2.12 Flowchart of the nipple localization algorithm. (a-b) Segmented fat (yellow), sternum (blue) and skin surface (red). (c) True positive candidate. (d) False positive candidate. (e-f) Determined nipple marked by square, with fitting plane shown in red and surface normal shown in blue. (a,c,d,e) are shown in axial view and (b,f) are shown in sagittal view.

Skin surface $\boldsymbol{S}$, fat tissue $\boldsymbol{F}$ and jugular notch (j_x, j_y, j_z) are first obtained using the methods presented in our previous work [49, 50]. A 3D Cartesian coordinate system is used in the following as illustrated in Figure 2.12(a): x, y, and z axes are oriented in left-right, posterior-anterior, and cranial-caudal direction respectively. The left and right nipples are separated based on the lateral location to the jugular notch.

The spatial location features, $f_x(\mathbf{s})$, $f_y(\mathbf{s})$, and $f_z(\mathbf{s})$, for a given skin pixel $\mathbf{s} = (s_x, s_y, s_z) \in \boldsymbol{S}$ are defined based on the jugular notch (j_x, j_y, j_z):

$$f_x(\mathbf{s}) = |s_x - j_x| / dx \tag{2.9}$$

$$f_y(\mathbf{s}) = (s_y - j_y) / dy \tag{2.10}$$

$$f_z(\mathbf{s}) = s_z - j_z \tag{2.11}$$

where dx and dy are the lateral width and anteroposterior width of the body respectively. The features defined in (2.9) and (2.10) are normalized by dx and dy to take into account of different size of the body. The absolute difference used in (2.9) allows the left and right breasts to be considered separately while using the same classifier trained with left and right nipples together.

For the feature computation of the depth of density and the degree of protuberance at a given skin pixel **s**, a 3D plane $\boldsymbol{P}(\mathbf{s})$ as defined in (2.12) and illustrated in Figure 2.12 (e-f) is first fitted in the sense of least squared error for a set of surface skin pixels $\boldsymbol{N}(\mathbf{s})$ in the neighborhood centered at **s** with radius of d_n mm.

$$(x - c_x, y - c_y, z - c_z) \bullet (n_x, n_y, n_z) = 0 \tag{2.12}$$

Where • is the inner product of two vectors; $\mathbf{n} = (n_x, n_y, n_z)$ is the surface normal vector as illustrated in Figure 2.12(e-f), and it is ensured to point towards the body; (c_x, c_y, c_z) is the centroid of the neighborhood set $\boldsymbol{N}(\mathbf{s})$, and it is ensured to be on the fitted plane.

The depth of density features $f_d(\mathbf{s})$ and $f_{\Delta d}(\mathbf{s})$ are then defined as follows:

$$f_d(\mathbf{s}) = \min\{\text{distance}(\mathbf{s}, \mathbf{f}), \text{ for } \mathbf{f} \in \boldsymbol{F}, (\mathbf{f} - \mathbf{s}) \bullet \mathbf{n} = |\mathbf{f}||\mathbf{n}|\} \tag{2.13}$$

$$f_{\Delta d}(\mathbf{s}) = f_d(\mathbf{s}) - \min\{f_d(\mathbf{t}), \text{ for } \mathbf{t} \in \boldsymbol{N}(\mathbf{s})\} \tag{2.14}$$

where $f_d(\mathbf{s})$ is the depth of density (non-fat tissue) along the surface normal direction **n** under skin pixel **s**; $f_{\Delta d}(\mathbf{s})$ is the difference between density depth at **s** and the local minimum depth in the neighborhood $\boldsymbol{N}(\mathbf{s})$.

The degree of protuberance $f_p(\mathbf{s})$ is a measure of the amount of deviation of **s** from the fitted plane $\boldsymbol{P}(\mathbf{s})$, which is the distance from **s** to $\boldsymbol{P}(\mathbf{s})$ as defined in (2.15):

$$f_p(\mathbf{s}) = (c_x - s_x, c_y - s_y, c_z - s_z) \bullet (n_x, n_y, n_z) \tag{2.15}$$

As the surface normal $\mathbf{n} = (n_x, n_y, n_z)$ is ensured to point towards the body, $f_p(\mathbf{s}) > 0$ at the "bump" pointing out from the body; while $f_p(\mathbf{s}) < 0$ at the "dent" pointing inwards to the body.

The set of features described above are computed for each skin surface pixel **s** that is inferior to the jugular notch and located in the anterior 2/3 of the skin surface. A skin surface

pixel **s** is considered to be a nipple candidate if either of the following two criteria is satisfied as illustrated in Figure 2.12 (b-c):

(1). $|f_p(\mathbf{s})|$ is a local maximum $\boldsymbol{N}(\mathbf{s})$ and $|f_p(\mathbf{s})| > t_p$.

(2). $f_d(\mathbf{s})$ is a local maximum in $\boldsymbol{N}(\mathbf{s})$, $f_d(\mathbf{s}) > t_d$ and $f_{\Delta d}(\mathbf{s}) > t_{\Delta d}$.

A machine learning classifier is employed to discriminate the true nipples from the false positive candidates, i.e., candidates that are proposed but are not nipples as shown in Figure 2.12(c). Seven commonly used classifiers, including logistical regression, SVM linear, SVM polynomial, SVM RBF, decision tree, random forest, and AdaBoosted tree, were explored. All features described above and the truth label (1 for candidate within 13 mm to the true nipple location; 0 for other candidates) for each proposed candidate in the training set are used to train the classifiers. The hyper-parameters of the classifiers are tuned based on the performance of the validation set.

During the testing phase, for each of the trained classifier and each breast in the testing set, all proposed candidates are first tested with the classifier, and a prediction score is obtained, where a higher score indicates a greater probability of being a nipple. The candidate with the highest prediction score for each side of the body, is then considered as the nipple location as illustrated in Figure 2.12 (d-e).

2.2 Quantitative imaging biomarkers from the breast

Two imaging biomarkers, breast density for female subjects and gynecomastia quantification for male subjects, are measured within the regions of interest, which are defined based on the segmented whole breast region and localized nipples.

2.2.1 Female breast density quantification

The region of interest used for the breast density measurement is denoted as local whole breast region and defined as the segmented breast region within d mm from the fibroglandular tissue (i.e., segmented soft tissues contained in the segmented breasts). The volume ratio of the fibroglandular tissue to the local whole breast region is then calculated and reported as the CT breast density. The left and right breasts are considered separately, and the higher density is used as the reported breast density as done in clinical practice [28].

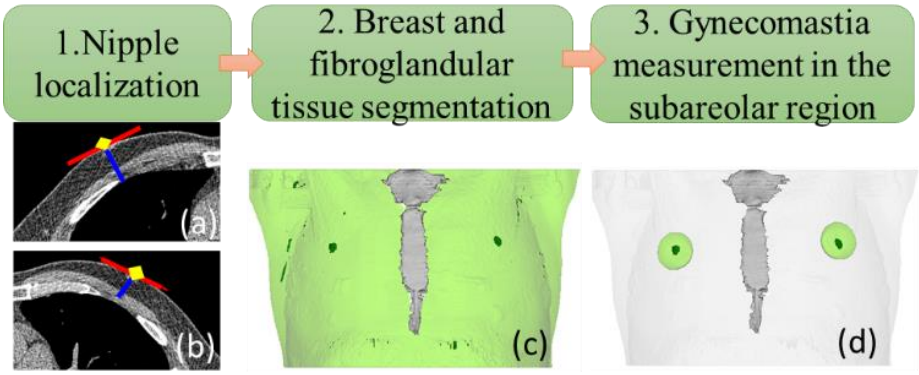

Figure 2.13. Flowchart of the gynecomastia detection framework. (a-b) Localized nipples marked by squares. (c) Segmentation of breast (light green) and fibroglandular tissue (solid green). (d) Identified subareolar region (light green) and gynecomastia (solid green). The segmented sternum (grey) is also shown in (c-d) for reference.

2.2.2 Male gynecomastia detection

The fully automated gynecomastia detection framework from LDCT consists of three main stages as illustrated in Figure 2.13. First, the nipple is localized for each side of the body. Second, the whole breast and the fibroglandular tissue is segmented based on the muscle front propagation algorithm presented in section 2.1. Third, the subareolar region is then defined as the segmented breast region within d_s mm to the localized nipple obtained using algorithm presented in section 2.2. The volume of segmented fibroglandular tissue within the subareolar region is measured for the gynecomastia assessment.

2.3 Experiments

The breast segmentation framework was validated using 1270 non-contrast LDCT images from ELCAP, LIDC and FAMRI datasets, most of which were performed for lung cancer screening. Visual inspection was used to evaluate the performance of the whole breast segmentation in both 2D axial view and coronal 3D view as shown in Figure 2.14. The results were considered unacceptable if segmentation errors, such as inclusion of muscle into the breast region or under-segmentation of fibroglandular tissue, may influence further breast analysis including density assessment and breast mass detection.

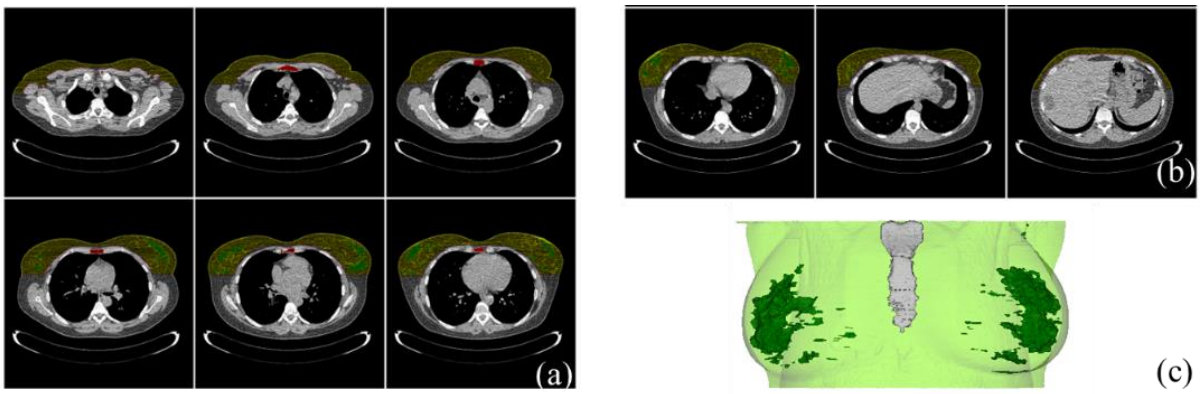

Figure 2.14. An example of visual inspection used during validation of breast segmentation. (a-b) 9 axial views uniformly sampled along the vertical range where segmented fibroglandular tissue (green), breast (yellow) and sternum (red) are shown. (c) 3D coronal view where the segmented fibroglandular tissue (dark green), breast (light green) and sternum (grey) is shown.

The nipple localization algorithm was validated using the training set and validation set consisting of 440 nipples (from 220 non-contrast LDCT scans) selected from ELCAP and LIDC datasets, where the gender information is unavailable. 85% (374) nipples were used for training and 15% (66) nipples were used for validation. The testing set consists of 838 nipples (from 419 non-contrast LDCT scans) selected from FAMRI dataset, consisting of 448 male

nipples and 390 female nipples. As these LDCT scans were acquired during lung cancer screening focusing on lungs, the breast region may be truncated in some scans. The dataset used here are constructed by including all scans having both nipples within the scan field of view from the three datasets. The ground truth nipple location was annotated manually for each breast. The outcome of the nipple localization algorithm is considered correct if the distance between the predicted nipple and the ground truth is within 2 cm. 2 cm is employed because it is approximately the maximal diameter of the nipples in the training set.

The automated breast density assessment was validated using a subset of 100 scans of female subjects randomly selected from the validation set. The density ground truth was established by an experienced radiologist (L.M., 28 years of mammography experience and 23 years of CT experience) by reviewing the CT scans and classifying each case into one of the 4 categories following the latest BI-RADS guidelines [28]: grade ‘a’ for breasts that are almost entirely fatty; grade ‘b’ for breasts with scattered areas of fibroglandular density; grade ‘c’ for breasts that are heterogeneously dense; and grade ‘d’ for breasts that are extremely dense. Cases located at the borderline between two categories were also marked by the radiologist. The distribution of reference density assigned by the radiologist is shown in Figure 2.15, which generally agrees with the common distribution across density categories as reported in the BI-RADS guidelines [28].

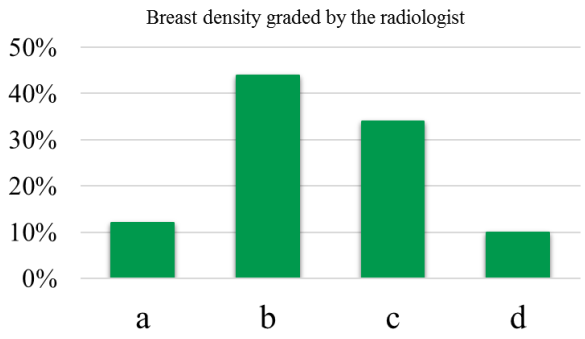

Figure 2.15. The distribution of the BI-RADS breast density assigned by the radiologist for the 100 scans used for the validation of the automated breast density assessment.

The gynecomastia quantification framework was validated using 448 breasts from LDCT scans of 224 adult men, which is the male subset of the testing set used in the evaluation of nipple localization. As there is still no consensus on the standard measure for the diagnose and quantification of gynecomastia [34], the gynecomastia reference standard used in this paper was established by an experienced radiologist who specializes in breast imaging using a 5-category grading scheme, where left and right breasts are considered separately. Grade 0, 0.8, 1, 2, 3 are assigned respectively to breasts with increasing amount of fibroglandular tissue in the subareolar region following the below guideline:

Grade 0: for breasts that are almost entirely fatty in the subareolar region.

Grade 0.8: for breasts with minimal fibroglandular density in the subareolar region.

Grade 1: for breasts with mild fibroglandular density (axial diameter < 2 cm) in the subareolar region.

Grade 2: for breasts with moderate fibroglandular density (axial diameter between 2 cm and 4 cm) in the subareolar region.

Grade 3: for breasts with marked fibroglandular density (axial diameter > 4 cm) in the subareolar region.

Examples of breasts of different reference grades are shown in Figure 2.16. Grade 0 – 3 account for 33.0%, 32.8%, 21.4%, 9% and 3.8% of the whole dataset respectively.

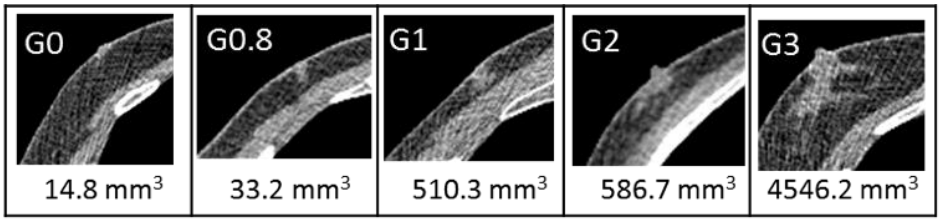

Figure 2.16 Examples of male breasts of different gynecomastia reference grades. For each figure, the reference grade is shown on the top and the automated measurement is shown on the bottom.

The Spearman's rank correlation was used to measure the concordance between the continuous gynecomastia measurements obtained by the automated framework and the categorical reference grades. To evaluate the performance for gynecomastia detection, breasts with reference grades of 0 and 0.8 were considered as negative and breasts with reference grades of 1 and above were considered as positive. The Receiver Operating Characteristic (ROC) curve and its area under the curve (AUC) were used to evaluate the diagnostic performance of the automated gynecomastia measurements.

2.4 Results

Satisfactory breast segmentation was achieved in 96.1% of the total 1270 LDCT scans, and only 0.79% of the scans had unacceptable segmentation due to the breast segmentation algorithm. 80% of the unacceptable segmentation was caused by the failure of prior dependence algorithms. Four examples of segmentation arranged in the order of increasing breast density are shown in Figure 2.17.

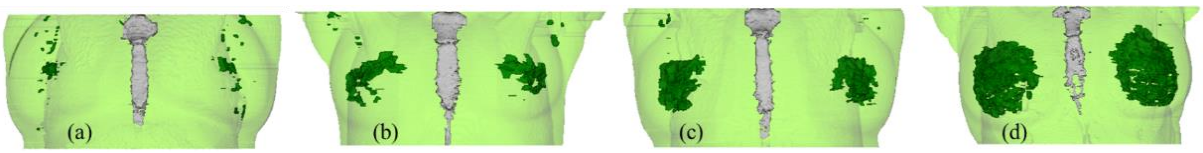

Figure 2.17. (a-d) Four cases in 3D coronal view in the order of increasing breast density where the segmented breast (light green), fibroglandular tissue (dark green) and sternum (grey) are shown.

All true nipples have been identified during the candidate proposal phase with on average 11.8 (median of 9) false positive candidates per breast. False positive nipple candidates as shown in Figure 2.12 (d) usually occur at skin folds, moles or uneven skin depth. The performance of using different machine learning classifiers are very similar. The best performance is obtained by using an SVM polynomial with degree of 2: 99.2% (831 out of 838) nipples in the testing set were correctly localized. 7 nipples consisting of 1 male nipple and 6 female nipples, were incorrectly localized. The incorrectly localized male nipple is shown in Figure 2.4 (b), where either the density depth beneath the skin or the degree of protuberance on the skin surface is insufficient. The 7 incorrectly localized female nipples are all from remarkably dense breasts, where the density depth at the nipple is not necessarily located at the local maximum as illustrated in Figure 2.4 (d).

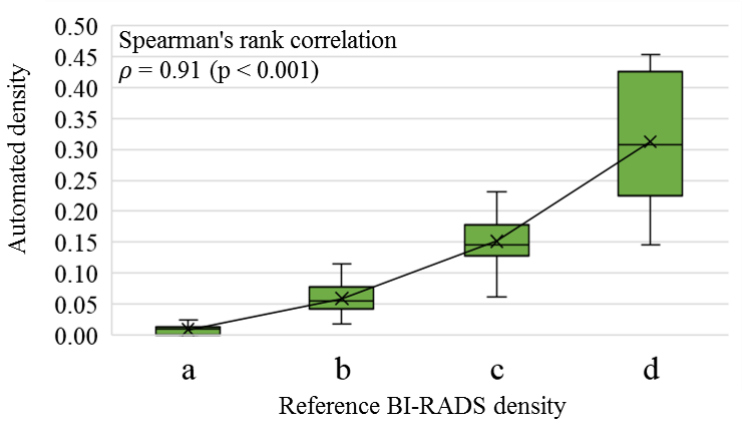

Figure 2.18. The box-and-whisker plot of the continuous density measurement obtained by the automated framework with respect to the subjective BIRADS density categories as the reference.

Table 2.1. The confusion matrix of the automated categorical density assessment and the reference BI-RADS density.

		Automated				
		1	2	3	4	total
Reference	a	12	0	0	0	12
	b	4	36	4	0	44
	c	0	1	33	0	34
	d	0	0	2	8	10

The comparison of the continuous breast density measurement and reference subjective grading is shown in Figure 2.18, where the Spearman's rank correlation was 0.91 (p-value < 0.001). After converting the automated measurements to categorical values using percentage cutoffs, the comparison is summarized by the confusion matrix shown in Table 2.1. Only 9 of the 100 scans were classified differently by the automated framework and the radiologist (by only 1 category difference), leading to the same density assessment in 91% cases. Since 6 out of these 9 scans were marked as at borderline by the radiologist, the breast density measurement of 97% scans can be considered consistent.

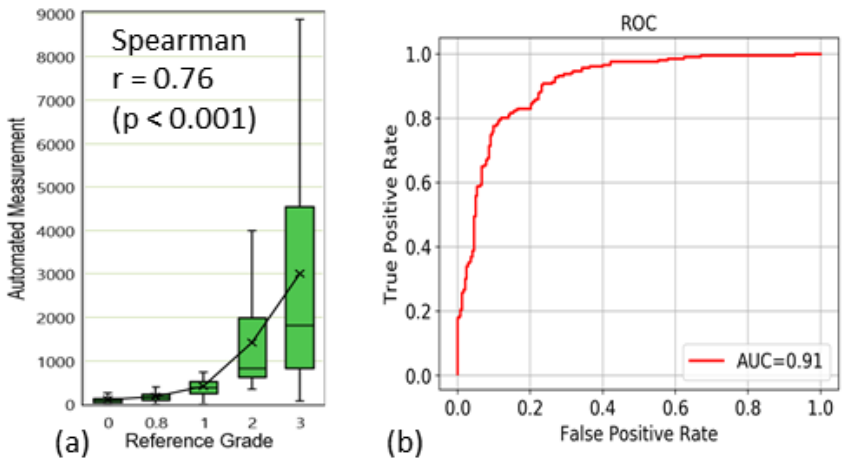

Figure 2.19. (a) Box-and-whisker plot of the automated gynecomastia measurements w.r.t. the reference grades. (b) ROC curves of using automated measurements for gynecomastia diagnosis.

For the evaluation of the nipple localization based gynecomastia detection, the Spearman correlation r = 0.76 (p-value < 0.001) is obtained between the automated gynecomastia measurements and the reference grades. The box-and-whisker plot of gynecomastia measurements is shown in Figure 2.19 (a). Examples of automated measurements for breasts of different reference grades are shown in Figure 2.16. The diagnostic performance of the automated gynecomastia measurements is given by the ROC curve shown in Figure 2.19 (b), in which the AUC = 0.91. Compared to the results presented in our previous study [52], the benefits of including the nipple location in the gynecomastia detection framework are given by the statistically significant improvement in Spearman correlation (0.70 vs 0.76, p-value < 0.001) and a positive trend in the AUC (0.86 vs 0.91, p-value = 0.065).

2.5 Discussion

The separation of pectoral muscles from the fibroglandular tissue is regarded as one of the greatest challenges for breast segmentation [46, 47] due to the lack of a well-define

boundary and the large individual variations of the fibroglandular tissues in terms of position, volume and shape as indicated in Figure 2.7. The solution presented in this study, muscle front propagation, focuses on the anterior front of the pectoral muscle that is much better-defined with a regular shape and smooth boundary.

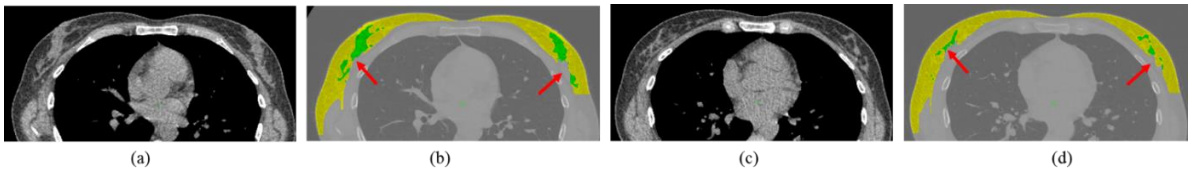

Figure 2.20. Example 1 of radiologist/computer discordance. (a,c) Two axial slices of the original CT scan. (b,d) The corresponding segmentation where the fibroglandular tissue (green), breast (yellow) and sternum (red) are shown. The arrows indicate the locations with over-segmentation of fibroglandular tissue.

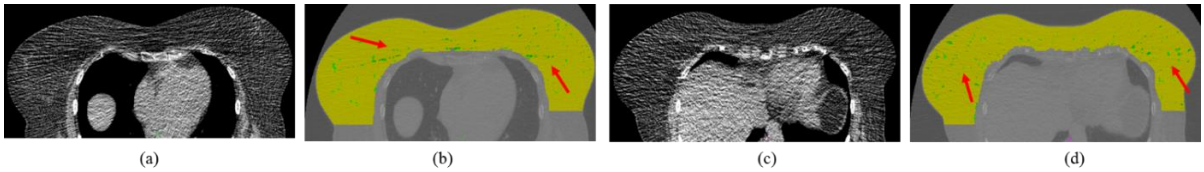

Figure 2.21. Example 2 of radiologist/computer discordance. (a,c) Two axial slices of the original CT scan. (b,d) The corresponding segmentation where the fibroglandular tissue (green), breast (yellow) and sternum (red) are shown. The arrows indicate the locations with over-segmentation of fibroglandular tissue due to high level of noise.

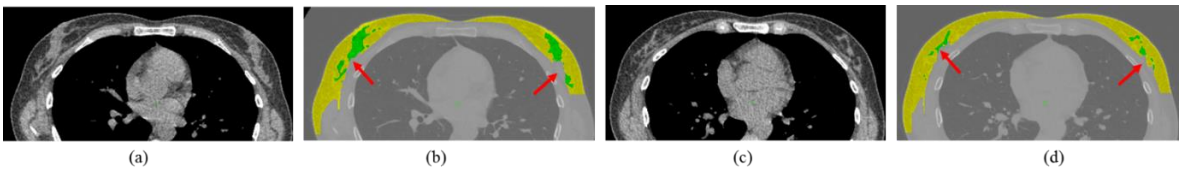

Figure 2.22. Example 3 of radiologist/computer discordance. (a,c) Two axial slices of the original CT scan. (b,d) The corresponding segmentation where the fibroglandular tissue (green), breast (yellow) and sternum (red) are shown. The arrows indicate the locations with under-segmentation of fibroglandular tissue.

The muscle front is modeled as a surface f (x,z) that is parameterized with image coordinates, instead of arc length as used by active contour models. It avoids the periodic re-parameterization and facilitates the front evolution, as for each x coordinate only variation in the y direction is considered, with the guarantee of a closed breast boundary consisting of the

anterior skin and the posterior muscle front on each axial level. This simplified contour model may not work well in a general contour propagating scenario, whereas it is sufficient for the muscle front due to its approximately convex shape and smoothness.

There are only 3 cases (out of 100) for which the density assessment by the automated framework is inconsistent with that of the radiologist, demonstrating the encouraging performance of the presented automated approaches. As shown in Figures 2.20-2.22, they all contain different types of segmentation issues, such as inclusion of muscles into the breast region as shown in Figure 2.20, under-segmentation of fibroglandular tissue as shown in Figure 2.21, and miss-classification of fibroglandular tissue due to high level of image noise as shown in Figure 2.22. For the future work, we plan to gather more discordant cases or cases with apparent problems and we plan to develop a more advanced density measurement approach, instead of the simple percentage volume ratio used here, so that the algorithm is more robust with respect to subtle segmentation issues.

The presented study is one of the first few studies [44, 48] on automated nipple localization from 3D imaging modality, which can potentially aid the aid a number of automated breast analysis tasks, such as breast density assessment [43], breast mass detection [41], breast cancer diagnosis [41, 42] and gynecomastia detection. The nipple localization algorithm was developed and evaluated for both female and male subjects, whereas the prior work on the automated breast analysis mainly focus on female subjects. 99.2% (831 out of 838) nipples in the testing set were correctly localized, which demonstrates the strength of the proposed anatomy-orientated image features and the efficacy of the machine learning classifiers on the task of fully automated nipple localization from LDCT.

Except our own prior work [52], only a small number of studies [34, 35] have investigated gynecomastia on CT scans, and these are all based on subjective reading by radiologists. Sonnenblick et al have demonstrated the feasibility of using chest CT for the confirmation of diagnosis for men with clinical symptoms of gynecomastia [34]. The results presented in this study further shows the feasibility of gynecomastia detection from LDCT using fully automated system, which may aid the early detection as well as the treatment of both gynecomastia and the underlying medical problems, if any, that cause gynecomastia.

2.6 Conclusion

A fully automated framework has been presented for breast heathy analysis from the LDCT acquired during annual lung cancer screening. The whole breast region is first segmented using an anatomy-orientated approach. The subareolar region is then localized using a machine learning based nipple detection algorithm. Finally, building upon the region of interest defined by the segmented breast and nipple location, two quantitative image biomarker measurements, the breast density for women and gynecomastia quantification for men, are accomplished. The automated breast segmentation has been visually validated with 1270 LDCT scans and achieved satisfactory outcomes in 96.1% scans. The automated density assessment was consistent with subjective reading of an experienced radiologist in 97 of 100 scans. Breast density assessment from LDCT can potentially serve as a valuable resource providing useful information with respect to breast cancer risk evaluation for many women who have undergone LDCT but not recent mammograms. The automated gynecomastia quantification was validation using 454 breast regions from non-contrast LDCT scans of 227 adult men. The gynecomastia reference standard was established by an experienced radiologist by reviewing the CT scans and classifying each breast into one of the five

categorical scores. The automated gynecomastia measurements have been demonstrated to achieve promising performance for the gynecomastia diagnosis with the AUC of 0.91 for the ROC curve and have statistically significant Spearman correlation $r=0.76$ ($p < 0.001$) with the reference categorical grades. The encouraging results demonstrate the feasibility of fully automated gynecomastia quantification from LDCT, which may aid the early detection as well as the treatment of both gynecomastia and the underlying medical problems, if any, that cause gynecomastia.

CHAPTER 3

FULLY AUTOMATED BONE ANALYSIS AND QUANTITATIVE BIOMARKER MEASUREMENTS

Skeletal structures, including clavicles, sternum, ribs and vertebrae that are covered by low-dose chest CT (LDCT), establish a reliable frame of reference to other non-rigid human organs in the chest region. Therefore, the segmentation and labeling of individual bones usually serve as a necessary prior step to the automated analysis of other organs, such as breasts [46, 45, 53], heart [54, 55, 56, 57] and lungs [58, 59]. In addition, quantitative image biomarkers, such as bone mineral density [60], vertebra compression fracture [61] and spinal curvature [62], measured from the spine provide valuable information for the diagnosis and treatment of a variety of skeletal diseases, such as osteoporosis and bone deformity.

The main contribution of the study is to present a fully automated framework for the bone mineral density quantification from LDCT scans, which has never been addressed in any previous publications, and provides the opportunity for the concurrent osteoporosis screening with annual lung cancer screening. The framework consists of two main stages: First, individual bone structures, including clavicles, sternum, ribs and vertebrae, are segmented and labeled with anatomical names, as presented in section 3.1. Second, individual vertebral bodies are individually segmented and the bone mineral density (BMD) is quantified as presented in section 3.2.

3.1 Individual bone structure segmentation and labeling from low-dose chest CT

The segmentation and labeling of the individual bone structures serve as the first step to the fully automated measurement of skeletal characteristics and the detection of

abnormalities. For example, spinal curvature has been used for the diagnosis of skeletal deformities [62]; bone mineral density measurement has been used for the diagnosis of osteoporosis [60]; quantitative geometry and intensity analysis of the individual vertebra has been used for the detection of vertebral fractures [61]. Moreover, the landmarks identified on the respective bone structures can potentially provide relatively robust and reliable location reference to other non-rigid human organs, such as breast [46, 45, 53], heart [56, 55, 54, 57] and lung [58, 59] that may have variable position and shape across individuals and during breathing, thereby facilitating the corresponding segmentation and analysis. Furthermore, these skeletal landmarks can also be employed to establish a standard frame of reference [58] in the chest region that allows image registration of cross-sectional CT scans as well as longitudinal scans of the same subject through non-linear transformations.

A fully automated framework for the segmentation and labeling of individual bone structures from LDCT is investigated in this study. These skeletal structures include both clavicles, sternum, all ribs and thoracic vertebrae. In addition, 28 robust skeletal landmarks from these structures are also identified. The proposed anatomy-directed system consists of four main stages as outlined in the flowchart in Figure 3.1. The stages are ordered so that the subsequent stages are built upon the results of previous stages to optimize the overall success rate.

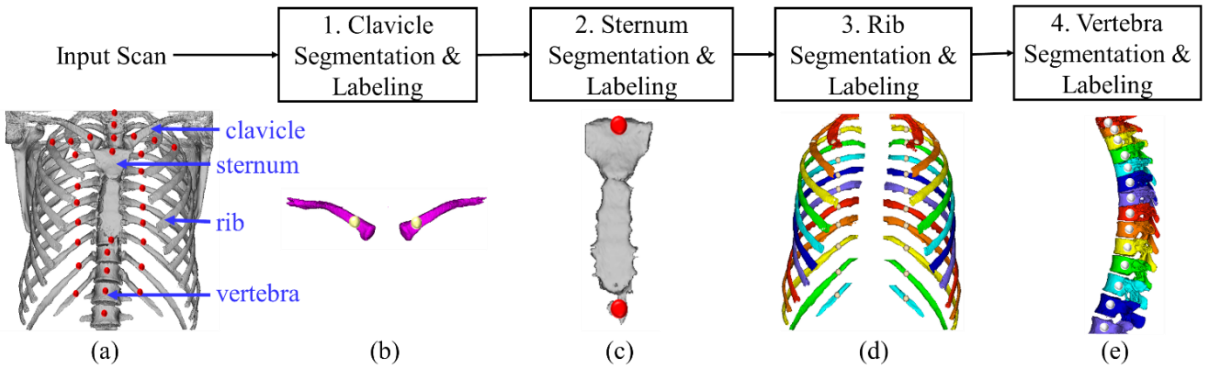

Figure 3.1. The flowchart of the bone segmentation and labeling framework. (a) The coronal visualization of the input scan generated by image thresholding and noise reduction filtering. (b-e) The segmentation and labeling of clavicles, sternum, ribs, and vertebrae. Note that left and right side are assigned with different labels but are shown in the same color here. The landmarks output in each stage are shown as dots.

This study addresses the bone segmentation and labeling in the context of the high image noise of chest LDCT. There have been a number of previous studies that focused on the segmentation of separate bone categories including: sternum by Liu et al [50], ribs by Lee et al [51], and vertebra by Reeves et al [63]. A study on the overall bone structure segmentation and recognition system from torso CT scans using intensity-based and anatomy-directed approaches has been presented by Zhou et al [64]. The segmented vertebrae, ribs, sternum and bones of upper limbs and lower limbs were evaluated with 48 torso scans and achieved above 90% success rate. Moreover, a model-based approach has been presented by Klinder et al [65] to segment and label individual ribs and vertebrae from chest CT scans. The mean rib cage model was applied to 18 CT scans (no CT protocol specified), resulting in successful segmentation and labeling in 16 scans. Additionally, significant effort has been devoted to the automated analysis of ribs [66] and vertebras [60, 61], however, these algorithms may not in general translate well to the recent chest LDCT protocols due to the higher levels of image noise and artifacts [49].

3.1.1 Method

The proposed framework consists of four main stages as outlined in Figure 3.1, where stage 2 and 3 are updates of previous published algorithms [50, 51]:

1) The clavicles are segmented by fitting a piecewise cylindrical envelope for the posterior part and by region growing for the anterior part. Two clavicles are then labeled as the right and left clavicles.

2) The sternum is segmented based on image intensity analysis under the spatial constraints provided by the segmented clavicles.

3) The individual ribs are segmented with anatomical names by 3D region growing within the volume of interest defined with reference to the spinal canal centerline and lungs. 12 right rib labels and 12 left rib labels are assigned to 24 ribs based on the spatial relationship.

4) The individual vertebrae are segmented based on image intensity analysis in the spatial region constrained by the previously segmented bone structures. Individual vertebra is labeled with its anatomical name, such as C7 (7^{th} cervical vertebra), T1(first thoracic vertebra), and L1 (first lumbar vertebra), by matching the vertebra and the labeled left and right ribs. Rib labels may also be corrected if unmatched left and right ribs connected to the same vertebra are detected.

3.1.1.1 Clavicle segmentation and labeling

The clavicle is a long bone that runs transversely and articulates medially with the manubrium of the sternum and laterally with the acromion of the scapula as illustrated in

Figure 3.1 (a). Located directly above the rib cage, it is usually the most superior bone structure visible in the thoracic CT scan, thereby serving as an important anatomical reference for the automated analysis and quantitative measurement of the chest CT. Accurate and robust segmentation of the clavicles provide a reliable starting point for the rib and sternum segmentation and labeling.

The clavicle is the most superior bone structure that is located at the same level with the airway along the anteroposterior axis. Therefore, a seed point on each of side of the clavicles can first be identified with respect to the segmented airway and serve as the starting point for the later segmentation. The posterior clavicles, defined as the clavicles posterior to the seeds, have well-defined shape that can be modeled by a piecewise cylindrical envelope; whereas the anterior clavicles, defined as the clavicles anterior to the seeds, are more difficult to model using a regular shape constraint. As a result, posterior clavicles and anterior clavicles are segmented using different approaches. The clavicle segmentation algorithm is summarized in the following three steps.

Step 1: Seed point identification

A seed point of the left/right clavicle as shown in Figure 3.2 (a) is detected at the center of mass of the most superior high-intensity pixels (intensity threshold Tc) located to the left/right of the airway. The airway is segmented using a previously published algorithm [67] and also shown in Figure 3.2 (a).

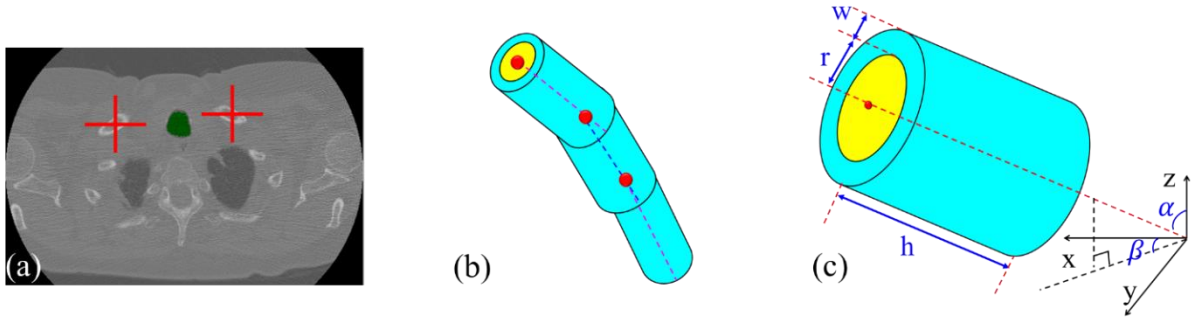

Figure 3.2. (a) The left/right seed (marked by the red cross) is located to the left/right of the airway (in green). (b) The piecewise cylindrical envelope is used to model the clavicle posterior to the seed. (c) Each cylinder segment consists of a central cylinder (yellow) and the outer shell (cyan).

Step 2: Segmentation of clavicles posterior to the seeds

The clavicle posterior to the seed point is modeled with a piecewise cylindrical envelope consisting of a series of cylinder segments with overlapping centerlines and variable radii and directions as indicated in Figure 3.2 (b). Each cylinder segment is composed of a central cylinder modeling the trabecular bone tissues of low image intensity (intensity threshold Tc) and an outer shell modeling the cortical bone of high image intensity as shown in Figure 3.2 (c). It is defined by 5 parameters as indicated in Figure 3.2 (c), including 2 pre-established parameters, the height h of the segment and the width w of the outer shell, and 3 parameters, the radius r of the central cylinder, and two angles α and β that determine the direction of the centerline, which are optimized after the fitting procedure.

Starting from the seed, a series of cylinders with overlapping centerlines are fitted into the scan iteratively by determining the optimal radius and direction of each cylinder segment that corresponds to the greatest matching score as defined in equation (3.1), which is the sum

of the percentage of low intensity pixels contained in the central cylinder and the percentage of high intensity pixels contained in the outer shell.

$$r, \alpha, \beta = \underset{r,\ \alpha,\ \beta}{\operatorname{argmax}} \sum_{(x,y,z)\ within\ the\ cylinder} \frac{1_{\text{central}}\,(\text{Im}(x,y,z)<\text{Tc}) + 1_{\text{shell}}\,(\text{Im}(x,y,z)>\text{Tc})}{\pi(r+w)^2\ h} \quad (3.1)$$

where Im(x,y,z) is the image intensity of pixel (x,y,z); $1_{central}$ (Im(x,y,z) < Tc) is the indicator function for pixels within the central cylinder and of low image intensity; 1_{shell} (Im(x,y,z) > Tc) is the indicator function for pixels within the outer shell and of high image intensity .

Step 3: Segmentation of clavicles anterior to the seeds

The clavicle anterior to the seed is segmented by successive region growing in the inferior direction in the spatial region constrained by the segmented clavicle cross section on the superior axial level. For a pixel to be considered as clavicle, two inclusion criteria are employed during the region growing process:

(i). It is a high-intensity pixel (intensity threshold Tc).

(ii). It is connected to the segmented clavicle cross section on the superior axial level.

The left and right seed points identified in step 1 are considered as two clavicle landmarks and denoted by L_{CL} and L_{CR} respectively as shown in Figure 3.1.

3.1.1.2 Sternum segmentation and labeling

The sternum is located in the medial and anterior thoracic cavity, and consists of three main parts as shown in Figure 3.3 (a): the manubrium, the sternal body and the xiphoid process. As sternum is connected to the clavicles at sternoclavicular joint as shown in Figure 3.1, a seed axial cross section of the sternum can first be obtained by identifying the most

superior high intensity pixels (intensity threshold Ts) that are medial and anterior to the clavicle landmarks L_{CL} and L_{CR} as shown in Figure 3.3 (b). Then the same region growing approach as described in section 3.1.1.1 is applied to segment cross sections in the superior and inferior direction respectively, starting from the seed axial cross section.

The existence of calcified cartilages is typically considered as the greatest challenge for the sternum segmentation [50] as they can be of the similar image intensity and closely connected to the sternum as shown in Figure 3.3 (c-d). The sternum segmentation algorithm takes into consideration that the cartilages only join the sternum in the costal space while there is usually well-defined sternum axial cross section in the intercostal space as shown in Figure 3.3 (a, c-e). During the successive region growing in the inferior direction, the lateral spatial limit of the sternum cross section can be defined based on the segmented sternum cross section on the superior axial level. If on current axial level, the segmented cross section exceeds the lateral limits, it suggests the existence of calcified cartilages and unreliable segmented cross section on current axial level (costal space). In such case, only high intensity pixels within the projection of previously segmented cross section are considered as sternum.

The current sternum segmentation algorithm is based on our previous published work [50], which was originally evaluated in 351 scans and demonstrated some segmentation issues when extending to the larger validation set of 1270 scans. The new algorithm takes advantage of the segmented clavicles to provide a reliable spatial reference to the superior end of the sternum and the lateral constraints to avoid the inclusion of adjacent calcified cartilages. The superior and inferior extent of the segmented sternum are considered as two landmarks and denoted by L_{SS} and L_{SI} respectively as shown in Figure 3.1.

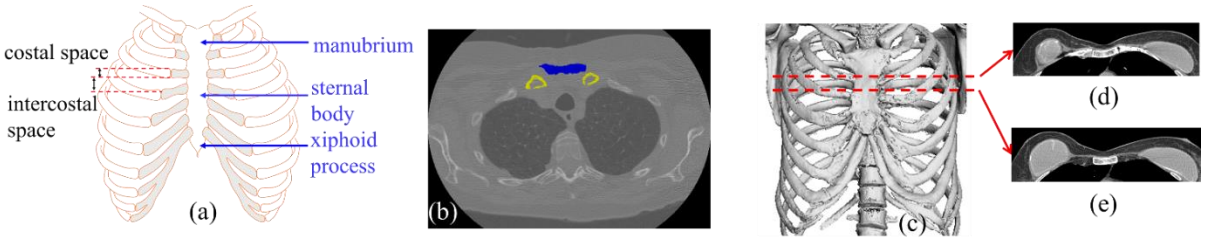

Figure 3.3 (a) Coronal view of the sternum and ribs. The costal cartilages are shown in grey. (b) The clavicle (yellow) and seed axial cross section of the sternum (blue). (c-e) A example scan with significant calcified cartilages shown in (c) coronal view, in (d) axial view at the level of costal space, and in (e) axial view at the level of intercostal space.

3.1.1.3 Rib segmentation and labeling

The 24 individual ribs are segmented by first detecting a seed point for each rib; and then by 3D region growing starting from the detected seed within the constrained region adjacent to lungs. The rib segmentation algorithm is built upon our previous work [51], where the seeds are detected in the region lateral to the spinal canal. As noted in [51], that algorithm may fail to detect superior ribs either because the scan does not cover up to that superior level or the scan are extremely noisy at that axial level due to the existence of the shoulder bones. To address these issues, in the new approach, the seeds of the superior two ribs on each side of the body are detected at the centers of masses of the most superior bone components located near the anterior surface of the lungs as illustrated in Figure 3.4. The new seeds are inferior to those used in the original algorithm, thereby addressing the issues of scan coverage and high noise levels. The seed of each individual rib is considered as a rib landmark and denoted as L_{RRN} and L_{LRN} for the n-th right and left rib respectively.

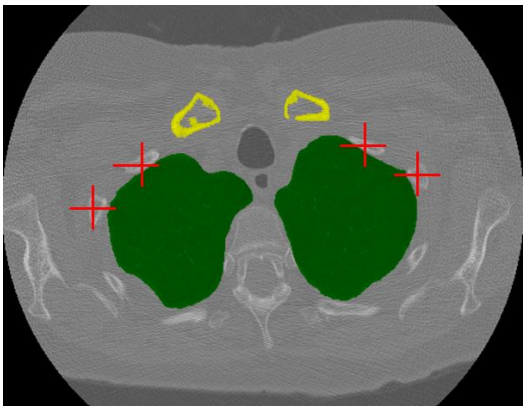

Figure 3.4. The seeds (marked by the red crosses) for the right/left first and second ribs. The clavicle (yellow) and lungs (green) are also shown.

3.1.1.4 Vertebra segmentation and labeling

Vertebrae are bones with irregular shape and complex structures, which articulate with each other and form the spinal column in the back of the human body. A typical vertebra consists of the vertebral body and the vertebral arch, which together enclose the spinal canal as illustrated in Figure 3.5 (a).

With the prior knowledge of the segmented sternum, ribs and lungs, the whole spine can be obtained based on image thresholding (at intensity Tv) and noise reduction filtering in the volume of interest that extends laterally to the ribs and lungs, and anteriorly to the sternum. Two consecutive vertebrae are subsequently separated from each other by fitting two dividing planes as shown in Figure 3.5 (b-d): one 3-layer plane anterior to the spinal canal for dividing the vertebral bodies; and the other single-layer plane posterior to the spinal canal for dividing the vertebral arches.

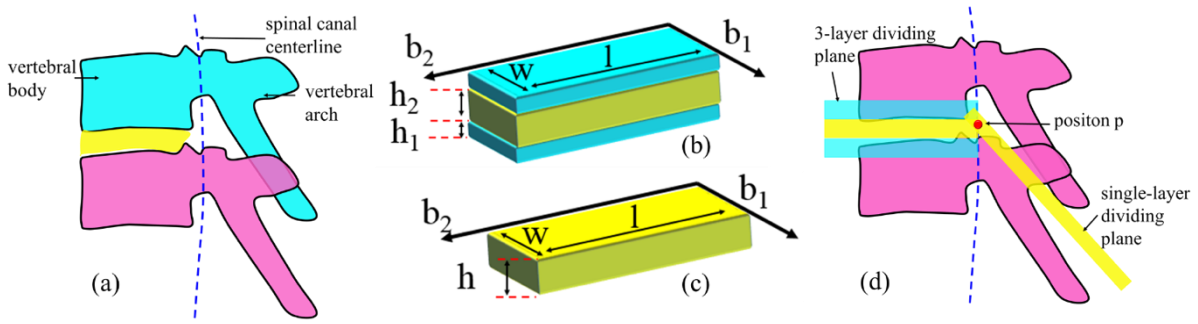

Figure 3.5. (a) Two consecutive vertebrae and the intervertebral disc (yellow) between them. (b) The 3-layer dividing plane model. (c) The single-layer dividing plane model. (d) The 3-layer plane model is fit in the region anterior to the spinal canal centerline and the single-layer plane model is fit in the region posterior to the spinal canal centerline.

The model of the 3-layer anterior dividing plane consists of the middle layer representing low intensity (intensity threshold at Tv) intervertebral disc and the two layers on top and bottom representing the high intensity endplates of the consecutive vertebrae as shown in Figure 3.5 (b, d). The dimensions of the plane are defined by 4 pre-established parameters including the height h_1 and h_2, width w and length l as indicated in Figure 3.5 (b). The position and orientation of the plane are determined by optimizing 3 parameters as shown in Figure 3.5 (b, d), including the vertical position p, and two orthogonal bases b_1 and b_2 by a fitting procedure based on equation (3.2).

$$p, b_1, b_2 = \underset{p,\ b1,\ b2}{\operatorname{argmax}} \sum_{\substack{(x,y,z)\ within \\ the\ plane}} \frac{1_{\text{disc}}\,(\text{Im}(x,y,z)<\text{Tv}) + 1_{\text{endplate}}\,(\text{Im}(x,y,z)>\text{Tv})}{(2h1+h2)wl} \qquad (3.2)$$

where the matching score is defined as the sum of the percentage of low-intensity pixels contained in the disc layer and the percentage of high-intensity pixels contained in endplate layers. Similar in equation (3.1), 1_{disc} (:) and 1_{endplate} (:) are indicator functions for pixels in the disc layer and endplay layers respectively.

The single-layer posterior dividing plane is used to model the low intensity soft tissue between two consecutive vertebral arches. Similar to the 3-layer model, it is defined by 6 parameters, including 3 pre-established parameters as specified in Figure 3.5 (c), height h, width w, and length l, for the dimensions, and 3 other parameters, p, b_1 and b_2, to optimize based on equation (3.3) for the determination of the position and orientation.

$$p, b_1, b_2 = \underset{p,\ b1,\ b2}{\operatorname{argmax}} \sum_{(x,y,z)\ within\ the\ plane} \frac{1(\mathrm{Im}(x,y,z)<\mathrm{Tv})}{hwl} \quad (3.3)$$

The segmented thoracic vertebrae are labeled with their anatomical names, thoracic vertebra 1, …, thoracic vertebra 12, based on their spatial relationships with the labeled ribs. The centers of mass of the 12 thoracic vertebrae are considered as 12 vertebra landmarks and denoted as L_{V1}, L_{V2}, …, LV_{12}. The cervical vertebrae and lumbar vertebrae that are superior and inferior the thoracic vertebrae are also labeled accordingly. The number of cervical vertebrae and lumbar vertebrae varies depending on the scan vertical range.

3.1.2 Experiments

The bone segmentation and labeling framework was validated using 1270 non-contrast LDCT chest scans from ELCAP [68], LIDC [69] and FAMRI datasets that were acquired generally for the purpose of lung cancer screening. The Validation by Visual Evaluation and Quantitative Revision (VEQR) [63, 4] was employed to evaluate the performance based on both 2D and 3D visualizations as shown in Figure 3.6-3.9.

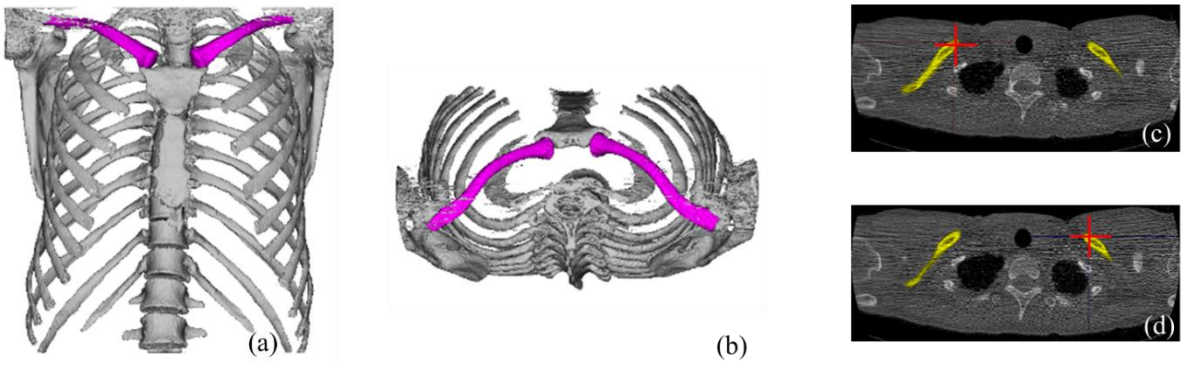

Figure 3.6. Visualizations used for the evaluation of the segmentation and labeling for clavicles. (a, b) Coronal and axial views for the clavicles (pink) and reference bones (gray). (c, d) Axial slices for the right and left landmarks (red crosses).

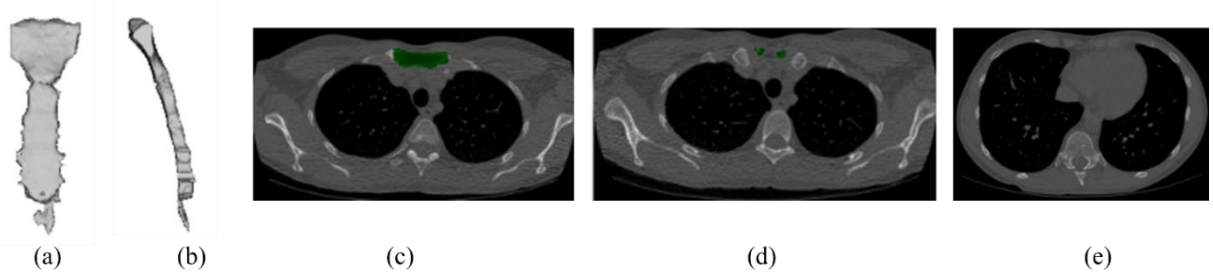

Figure 3.7. Visualizations used for the evaluation of the segmentation for sternum. (a, b) Coronal and sagittal views for the sternum. (c, d, e) Axial slices for the seed, superior end and inferior end of the sternum (green).

For the clavicles, the segmentation is visualized in 3D coronal view and axial view respectively with the bone segmentation (obtained by image thresholding and noise reduction) also shown as reference in Figure 3.6 (a, b). The clavicle landmarks L_{CL} and L_{CR} for left and right clavicles are visualized in 2D axial view in Figure 3.6 (c) and (d) respectively. For the sternum, the segmentation is visualized in 3D coronal view and sagittal view respectively in Figure 3.7 (a, b). The seed point, and two sternum landmarks L_{SS} and L_{SI}, are visualized in 2D axial view respectively in Figure 3.7 (c-e). For the ribs, the segmentation and anatomical labels are visualized in 3D coronal view with 24 rib landmarks also shown in Figure 3.8 (a).

The superior end of the right and left rib segmentation are shown in 2D axial views in Figure 3.8 (b, c) for the further verification of the anatomical labeling. For the vertebrae, the segmentation and anatomical labeling are visualized in 3D coronal view and sagittal view respectively in Figure 3.9 (a, b). The rib labeling is also visualized in transparent in Figure 3.9 (a) to serve as a reference for the verification.

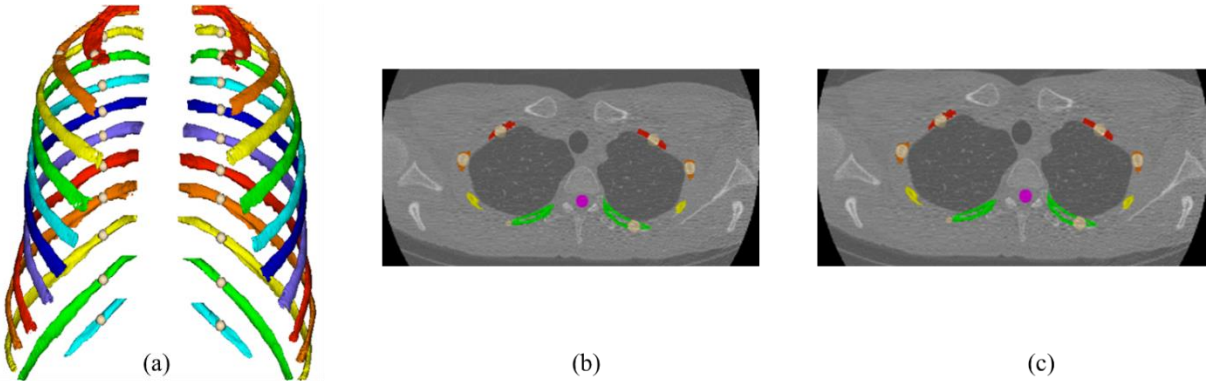

Figure 3.8. Visualizations used for the evaluation of the segmentation and labeling for ribs. (a) Coronal view for ribs and landmarks (white dots). (b, c) Axial slices for the right/left superior ribs.

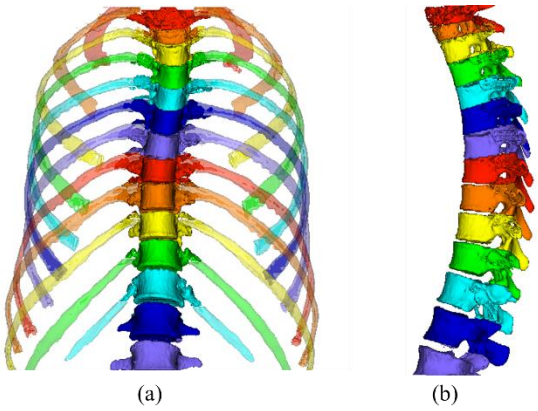

Figure 3.9. Visualizations used for the evaluation of the segmentation and labeling for vertebras. (a, b) Coronal and sagittal views for vertebrae. The labeled ribs are also shown in transparent colors as reference.

The results are considered as unacceptable if segmentation issues (including under-segmentation and over-segmentation) or labeling issues occur and may influence the

correctness of the corresponding landmark detection and the potential further analysis of the succeeding bone structures.

3.1.3 Results

The success rates from visual inspection for the bone structure segmentation and labeling framework in terms of clavicle, sternum, individual ribs and individual vertebrae segmentation and labeling are summarized in the Table 3.1.

Table 3.1 Success rates (of 1270 cases in total) for the bone structure segmentation and labeling framework.

	Segmentation	Labeling
Clavicles	97.1%	97.1%
Sternum	97.3%	
Ribs	97.2%	94.2%
Vertebra	92.4%	89.9%

3.1.4 Discussion

The success rates for the segmentation of individual bone structures were all above 90% as shown in Table 3.1. Especially for the segmentation in first three stages, the presented framework achieved satisfactory results in more than 97% cases, demonstrating the reliable and robust performance. Examples of unacceptable segmentation for clavicles, sternum and ribs are shown in Figure 3.10, 3.11, and 3.12 respectively. These illustrate two major causes for the unusable results in the first three stages: (1) the existence of metal implants as shown in Figure 3.10, which violates the assumptions about normal anatomy and has not been taken

into consideration by the current algorithm so far; (2) a high level of image noise leading to low local contrast of bone structures as shown in Figure 11-12, thus causing the under-segmentation of bones.

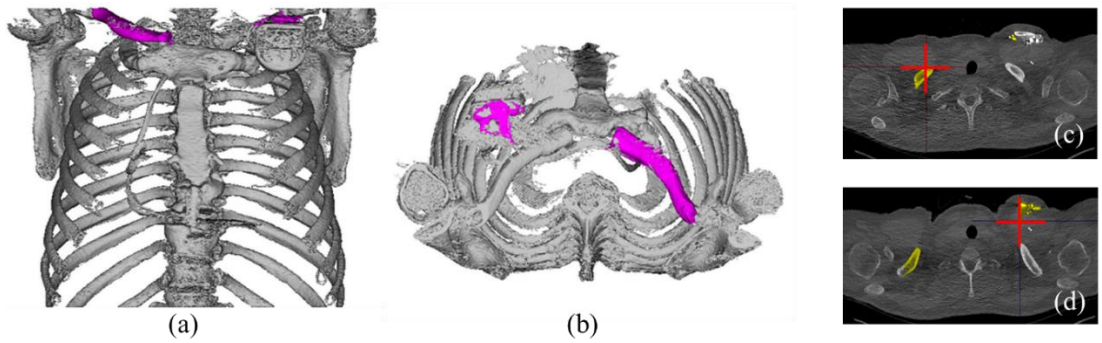

Figure 3.10. An example of unacceptable clavicle segmentation due to the existence of metal implant.

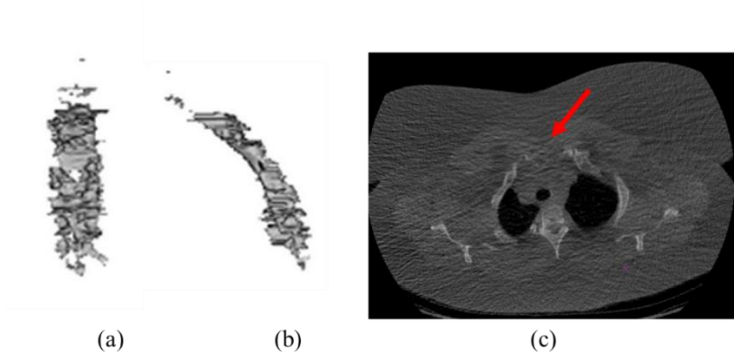

Figure 3.11. An example of unacceptable sternum segmentation due to high level of image noise.

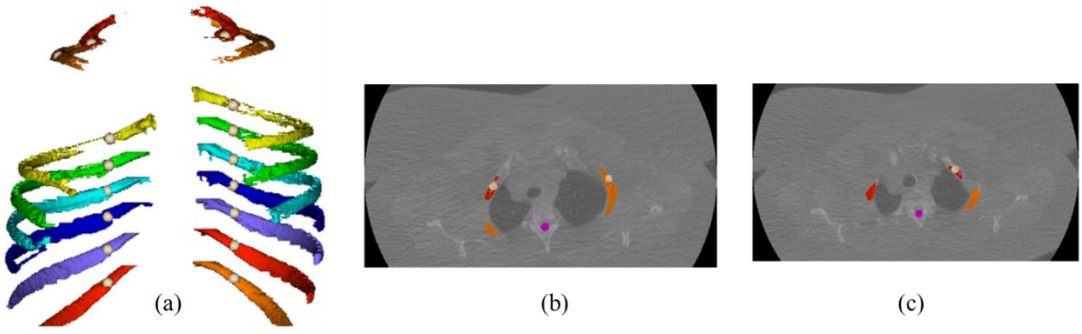

Figure 3.12. An example of unacceptable rib segmentation due to high level of image noise.

The performance of labeling is in general inferior to the performance of segmentation, since any segmentation issue will necessarily influence the consequent step of labeling, thus leading to a labeling issue. Moreover, a successful segmentation does not guarantee a successful anatomical labeling. The example shown in Figure 3.13 shows that, although the rib segmentation is successful as indicated in Figure 3.13 (a), the rib anatomical labeling is incorrect due to the assumption that the left n^{th} rib and right n^{th} rib are, in general, located at the same axial level (not true for the 4^{th} ribs in the example). Note that the performance for the clavicle segmentation and labeling is the same, because the clavicle labeling only involves determining left and right side which is trivial given a successful segmentation.

The success rate in the final stage, vertebra segmentation and labeling, is the worst because of two primary reasons: First, the final stage is built upon the previous stages; therefore, any segmentation or labeling issue occurring in the previous stages is accumulated in the final stage as shown in Figure 3.13, where the rib labeling issue leads to an unusable vertebra labeling although the vertebra segmentation is initially correct. Second, the whole spine is of variable curvature along the vertical direction and each vertebra consists of a number of anatomical components with irregular shapes as shown in Figure 3.5 (a), which makes it more challenging to model compared to the other three bone structures with relatively well-defined shape and fewer nearby structures of similar image intensities. The framework is arranged in the order of increasing difficulty so that the target structures segmented first can be used as spatial priors for the structures segmented later. An unacceptable example of vertebra segmentation and labeling is shown in Figure 3.14.

To resolve the issues discussed above, there are several possible directions of the future work. First, metal implants need to be taken into consideration to avoid being confused

with bone structures. Second, the image intensity related algorithm parameters need to be optimized for individual scan instead of being constant across all scans with different levels of noise and radiation dose, as well as various types of reconstruction methods, filtering kernels and CT scanners.

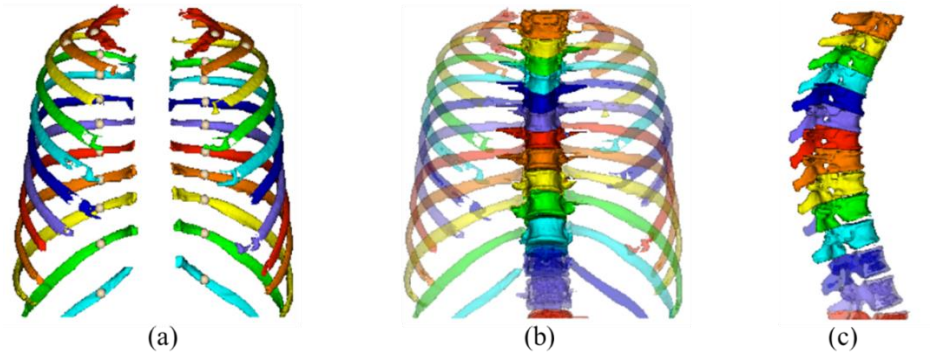

Figure 3.13. An example of acceptable (a) rib segmentation (a) but unacceptable (b) rib anatomical labeling, which also leads to unacceptable (b, c) vertebra labeling.

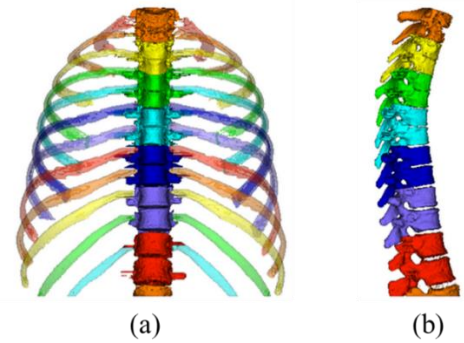

Figure 3.14. An example of unacceptable vertebra segmentation and labeling where the dividing planes are incorrect.

3.1.5 Conclusion

The automated segmentation and labeling of the individual bone structures are essential for the automated measurement of skeletal characteristics and the detection of abnormalities. It can potentially provide a relatively reliable location reference to other non-

rigid human organs and aid the image registration. A fully automated anatomy-directed framework for the segmentation and labeling of the individual bone structures including both clavicles, sternum, all ribs and thoracic vertebrae from LDCT is presented. The framework was validated with 1270 non-contrast LDCT images and visual evaluation results are encouraging for use in a fully automated image analysis context.

3.2 Bone mineral density quantification from low-dose chest CT

Osteoporosis is the most common metabolic bone disease and is estimated to affect 12.3 million US population aged 50 years or older in 2020 [18]. The prevalence of osteoporosis is continuing to increase rapidly with the progressively aging populations [20]. Osteoporosis is characterized by low bone density and micro-architectural deterioration of bone tissue [19] with related complications, such as osteoporotic fractures, becoming a significant cause of increased morbidity and mortality [20, 70], and thereby creating tremendous social and economic burdens. It is estimated by Burge et al. [71] that there are more than 2 million new osteoporotic fractures in the US every year, incurring more than $20 billion in costs. As osteoporosis is a silent disease and often undetected until a fracture occurs, the early diagnosis of osteoporosis is crucial for timely treatment and risk assessment for osteoporotic fractures [20, 70].

The assessment of bone mineral density (BMD) is central in the diagnosis and follow-up therapy monitoring of osteoporosis as well as other metabolic bone diseases [72]. Osteoporosis was defined by World Health Organization (WHO) [19] based on the measurement of areal bone mineral density (aBMD) (g/cm^2) at either the femoral neck or the lumbar spine using dual-energy X-ray absorptiometry (DXA), which is currently the most

widely used and gold standard technique for the diagnosis of osteoporosis and fracture risk estimation [72].

Computed tomography (CT) has also been applied to quantify BMD by providing separate volumetric bone mineral density (vBMD) measurements of trabecular and cortical bone, which is generally considered more accurate than two-dimensional DXA aBMD measurements that cannot distinguish between cortical and trabecular bone and can be distorted due to spinal degenerative changes (such as compression fracture, osteoarthritis, osteophytes and degrading vertebral disks), deformity or calcifications located near the spine such as aortic calcification [72, 70]. Both calibrated [73, 74, 75, 76, 77, 78, 79, 80, 81], and uncalibrated [73, 75, 77, 82, 83, 84, 85, 77, 86, 87] [88] vBMD obtained from CT have been shown to correlate well with aBMD obtained from DXA and be able to aid the diagnosis of osteoporosis and the detection of vertebral compression fracture. Calibrated vBMD can be obtained by using either an external reference phantom scanned with the patient [74, 73, 75, 76, 78, 89] or soft tissue (fat or muscle) of the patient as internal reference [77, 74, 76, 79] to derive the calibration equation to convert the CT attenuation in Hounsfield units (HU) to vBMD in milligrams of calcium hydroxyapatite per cubic centimeter (mg/cm^3). Uncalibrated vBMD expressed in HU has been demonstrated in several recent studies [73, 75, 77] to have no statistically different diagnostic performance compared to calibrated vBMD, suggesting the feasibility of using routine CT without calibration phantom for opportunistic BMD assessment and osteoporosis screening. As the chest LDCT in general also covers the regions of thoracic vertebrae T1-T12 and lumbar vertebrae L1-L2, it has the potential to serve as an opportunistic osteoporosis screening modality [21] as other routine CT obtained for clinical indications other than bone densitometry [87, 73, 75, 77, 80].

Previous studies [89, 73, 74, 75, 76, 77, 78, 82, 83, 84] [85, 77] on BMD assessment from CT mostly focus on measurement within 2D axial region of interest (ROI) placed manually in the trabecular region of the vertebral bodies. 3D volumetric ROI has been employed in recent studies [79, 80, 86, 88], which has been shown to provide more reliable measurements with greater reproducibility and correlation with aBMD measured from DXA (BMD_{DXA}) compared to 2D ROI [79, 86]. Tay et al demonstrated that using both trabecular and cortical bone can achieve better correlation with BMD_{DXA} in [79]. Fully automated BMD assessments from CT (BMD_{CT}) have been presented by Summers et al. for CT colonography [80], by Tay et al. for routine abdominal CT [79], by Zhou et al. for whole body CT [81], and by Burns et al. for CT with spine protocol [88].

The purpose of this study is to present a fully automated system for the BMD assessment based on image attenuation (HU) from the LDCT acquired during annual lung cancer screening using BMD_{DXA} measured from DXA as the reference standard. BMD_{CT} measurements of both trabecular and a combination of cortical and trabecular vertebral bodies at different vertebral levels were investigated by exploring various 3D volumetric region of interest (ROI) for the assessment. Our hypothesis is that BMD_{CT} can be obtained fully automatically from LDCT with statistically significant strong correlation with BMD_{DXA}, demonstrating the potential of opportunistic osteoporosis screening with concurrent lung cancer screening using LDCT.

3.2.1 Methods

The presented framework for BMD_{CT} assessment consists of three main stages as illustrated in Figure 3.15. First, the individual vertebra is segmented and labeled with its

anatomical name (n-th thoracic vertebra as Tn and n-th lumbar vertebra as Ln) using anatomical-directed knowledge-based approach that was presented in the previous section 3.1. Second, the vertebral body with closed 3D surface is then segmented using progressive surface resolution (PSR) algorithm by considering both image intensity and surface geometry. Third, a 3D volumetric region of interest (ROI) within each segmented vertebral body is determined for the mean CT attenuation measurement.

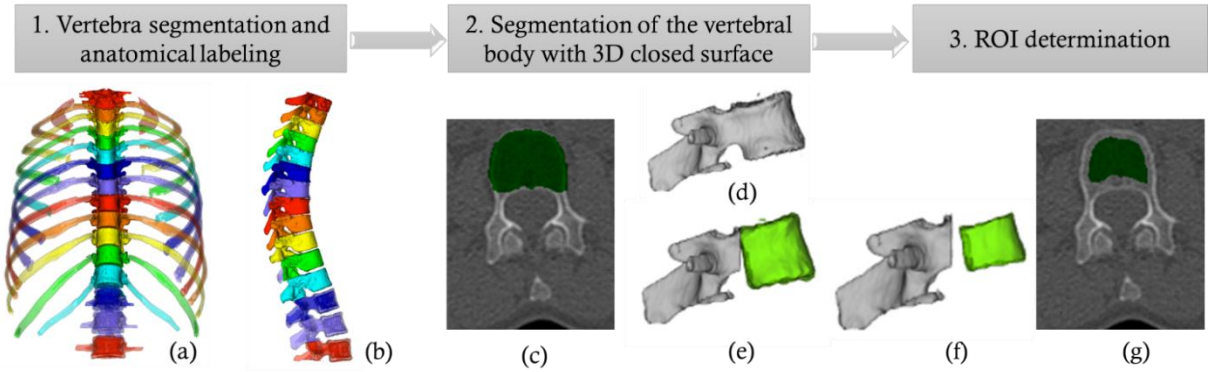

Figure 3.15. Flow chart illustration of the presented system for BMD_{CT} assessment. The segmentation and anatomical labeling of vertebrae in which the first three lumbar vertebrae are visible in (a) coronal view with ribs shown in transparent colors and in (b) sagittal view. The segmented individual vertebra and vertebral body (green) shown in (c) axial view and (d-e) sagittal view. The ROI (green) for BMD_{CT} measurements shown in (f) sagittal view and (g) axial view.

3.2.1.1 Segmentation of vertebral body with 3D closed surface

The segmentation of a vertebral body with 3D closed surface as shown in Figure 3.15 (c-e) is achieved using progressive surface resolution (PSR) algorithm that was previously presented in [90], which is a generic algorithm designed to segment the surface of approximately convex blob-like structures. The PSR algorithm requires an intensity image and a high confidence boundary image (which can be acquired automatically using simple methods such as 3D edge detection) as inputs. The target surface is realized by a closed

triangular mesh, which therefore guarantees the enclosure of the surface. The surface vertices on the triangular mesh are constrained along radial trajectories that are uniformly distributed in the 3D angle space, emanating from the centroid of the input high confidence boundary. The segmentation is accomplished by sequentially resolving each surface vertex on the respective radial trajectory according to a dynamic attraction map that incorporates the prior knowledge of the vertebral cortical surface regarding its intensity, the smoothness and the geometry.

The PSR algorithm consists of three main steps. First, trajectories are generated using rays emanating from the centroid of the input high confidence boundaries. Second, the surface is initialized based on intersections between the trajectories and the input high confidence boundaries. Third, sequential surface resolution is performed based on a dynamic attraction map, which describes the magnitude of the attraction potential $\mathbf{P}(p)$ that is intended to measure the attraction at an image point p to the target surface we seek to segment. The definition of attraction potential $\mathbf{P}(p)$ takes account of intensity, smoothness and shape constraints. It is dynamically updated by incorporating the location of newly resolved surface vertices, and then is used to resolve other unknown vertices. Each surface vertex is constrained along a unique radial trajectory and resolved in an order of decreasing degree of evidence regarding the target surface without involving any iterative refinement. This novel resolution strategy provides uniform angular resolution for the segmented surface with computation complexity and runtime that are linearly constrained by the total number of vertices on the triangular mesh.

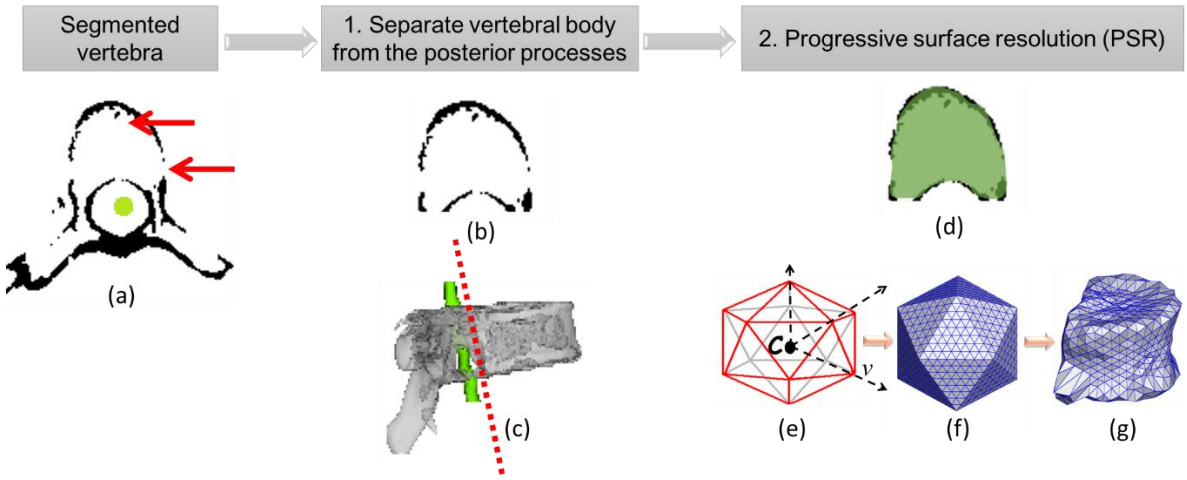

Figure 3.16. Flow chart illustration of the segmentation of vertebral body with 3D closed surface. (a) vertebra segmentation. (b) A high confidence boundary image ***B*** of the vertebral body as input to PSR algorithm. (c) Illustration of the separation between vertebral body and posterior part of the vertebra. (e-f) Initial surface generation. Resolved 3D closed surface of the vertebral body represented as triangular mesh (g) and green region (d) overlaid the input ***B***.

A high confidence boundary image ***B*** of the vertebral body is required as the input to the PSR algorithm to provide the centroid ***C*** and to initialize the surface, as illustrated in Figure 3.16 (b). ***B*** is obtained based on the following two steps. First, individual thoracic vertebra and the centerline of spinal canal are obtained following methods described in previous section 3.1, as illustrated in Figure 3.16 (a, c). Second, the vertebral body is separated from the spinal processes by analyzing the intensity profile along the anterior-posterior direction as illustrated in Figure 3.17. The boundary image ***B*** is then generated by removing the spinal processes from the vertebra segmentation as shown in Figure 3.16 (b, c), and the centroid ***C*** is determined automatically by computing the center of mass of ***B***. The input boundary image ***B*** may contain spurious high intensity components in the trabecular bone and holes on the cortical surface as illustrated in Figure 3.16 (b).

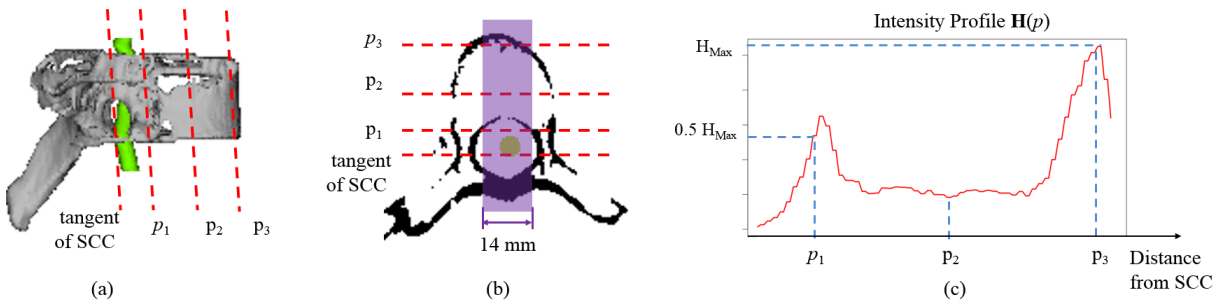

Figure 3.17. Determination of the dividing plane that separates the vertebral body and the spinal processes. The tangent of the SCC and three example plane candidates p_1, p_2, p_3 are shown in (a) sagittal view and (b) in axial view. Only the region of a lateral width of 14 mm are used to compute the intensity profile **H**(p) as illustrated by the shaded region in (b). (c) An example intensity profile.

The separation between the vertebral body and the spinal processes is accomplished by determining a dividing plane that is parallel to the tangent of the spinal canal centerline (SCC). An intensity profile **H**(p) is first defined as follows for any plane p that is anterior and parallel to the tangent of the SCC.

$$\mathbf{H}(p) = \sum_{v \in p,\ dist_x(v,\ SCC) \le 7\,mm} \boldsymbol{I}(v), \quad \forall\, p \in \boldsymbol{T} \tag{3.4}$$

Where **I** (v) is the interpolated image intensity of point v, which is sampled with resolution 0.5 mm x 0.5mm from plane p; $dist_x(v, SCC)$ is the lateral distance between v and the SCC; $\boldsymbol{T}$ is the set of all planes that are anterior and parallel to tangent of the SCC, where the separation between adjacent planes is 0.5 mm. Thus the intensity profile **H**(p) is computed by summing interpolated image intensities within a lateral width of 14 mm on each dividing plane candidate p as illustrated in Figure 3.17. The dividing plane p^* is then determined as follows:

$$p^* = \underset{p}{arg\ min}\ dist(p,\ SCC),\quad p \in \boldsymbol{T}\ and\ \boldsymbol{H}(p) > 0.5\, H_{Max} \tag{3.5}$$

Where dist(p, SCC) is the distance between the plane candidate p and the SCC, H_{Max} is the maxima of the intensity profile. Therefore, p^* is in fact the most posterior plane with sum intensity larger than 50% shoulder of the profile. An example intensity profile is shown in

Figure 3.17 (c), where several dividing plane candidates have been marked by dashed line, and p_1 is the selected dividing plane.

A closed triangular mesh is used to realize the vertebral surface and thereby guarantees the enclosure of the surface. The surface vertices on the triangular mesh are constrained along radial trajectories that are uniformly distributed in the 3D angle space. The trajectories are generated following two steps. First, the approximate centroid ***C*** of target structure is obtained by computing the center of mass of the input high confidence boundary ***B***. Second, project rays from ***C*** through vertices of a triangular tessellated icosahedron centered at ***C*** as shown in Figure 3.16 (e). Higher spatial resolution for the segmented surface is easily attained by subdividing each equilateral triangular face of the icosahedron into smaller equilateral triangular elements as shown in Figure 3.16 (f).

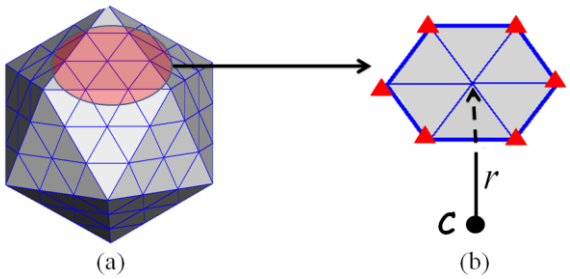

Figure 3.18. Neighbor set ***N***(*p*) of image point *p* located along trajectory *r*. The neighboring vertices of *p*, i.e., vertices in ***N***(*p*), are marked by triangles in (b).

The surface mesh is initialized based on the intersections between radial trajectories and input high confidence boundary ***B*** as summarized in Algorithm 1. For each trajectory intersecting ***B***, the location of the corresponding vertex is resolved according to the respective intersections, and considered to belong to the target surface with high confidence, with additional geometry constraints to reject intersections between trajectories and spurious

boundaries (i.e., boundaries representing high interior regions instead of the cortical surfaces in $\boldsymbol{B}$). The resolved set $\boldsymbol{S}$ is then initialized using these vertices. The vertices that are undermined yet constitute the unresolved set $\boldsymbol{F}$ (i.e., $\boldsymbol{S}^{C}$). The surface is then initialized by constructing the triangular mesh representation of vertices in $\boldsymbol{S}$ with respective neighboring relations defined by the trajectories. If $\boldsymbol{F} \neq \emptyset$, the initial surface is not close yet. However, after the subsequent surface resolution process, all vertices in $\boldsymbol{F}$ will be resolved in an ordered sequence, and a closed triangular mesh representation of the target surface will be achieved as shown in Figure 3.16 (g).

Algorithm 1 Surface initialization

```
/* Initialize the resolved set S and unresolved set F */
S ← { arg max dist(p, C), p ∈ r and p ∈ B | ∀ r ∈ R }
      p

F ← S^C


/* Reject intersections with low GNI or CI*/
done ← 0
while done = 0 do
   done ← 1

   for each v ∈ S do
        if GNI(v) < r1 or CI(v) < r2 do
              remove v from S, add v into F
            done ← 0
         end if
   end for

end while
```

The geometry constraints are introduced based on the assumptions that the target structure is nearly convex and the surface is smooth in the local neighborhood. We employ

two metrics, the good neighbor index (GNI) defined in (3.6) and the convexity index (CI) defined in (3.7), to reject intersections.

$$\mathrm{GNI}(v) = \frac{|\{v' | v' \in \boldsymbol{N}(v),\ v' \in \boldsymbol{S}\}|}{|\boldsymbol{N}(v)|}, \quad \text{for } v \in \boldsymbol{F} \tag{3.6}$$

Where | **A** | is the number of elements contained in set **A**; $\boldsymbol{N}(p)$ is the set of neighboring surface vertices of p. If we let r denote the trajectory associated with p, i.e. $p \in r$, $\boldsymbol{N}(p)$ consists of surface vertices located along neighboring trajectories, which are connected to r by edges on the surface mesh, as illustrated in Figure 3.18.

$$\mathrm{CI}(v) = \frac{\mathrm{dist}(v,\ \boldsymbol{C})}{\max\limits_{v' \in \boldsymbol{N}(v)} \mathrm{dist}(v',\ \boldsymbol{C})} \tag{3.7}$$

Where dist(v', $\boldsymbol{C}$) is the Euclidean distance between vertex v' and the centroid $\boldsymbol{C}$. Note that for an unresolved vertex v' (i.e., $v' \in \boldsymbol{F}$), its distance to $\boldsymbol{C}$ should assume the local maximal distance from nearby resolved vertices to $\boldsymbol{C}$, so that the geometry information is passed through the unresolved vertices to the neighboring resolved vertices. This can be ensured in the implementation by assigning a local maximal distance to any vertex newly added to $\boldsymbol{F}$.

As a consequence, an intersection v between the trajectory and $\boldsymbol{B}$ is rejected if most neighboring trajectories have no intersections (i.e., GNI(v) is low) or the distance from v to the centroid $\boldsymbol{C}$ is significantly shorter than that from neighboring vertices (i.e., CI(v) is low). A low GNI(v) indicates the region around v is noisy and unreliable, thus v should not be considered as part of the robust boundary; a low CI(v) indicates that v introduces severe concavity into the surface, which violates the convexity assumption about the target structure.

The surface segmentation is accomplished by sequentially resolving all vertices in the unresolved set $\boldsymbol{F}$ according to the dynamic attraction map. The vertices with the most

evidence regarding the target surface are resolved first and then are incorporated to update the attraction map for the resolution of the remainder of the vertices in $\boldsymbol{F}$.

The dynamic attraction map describes the magnitude of the attraction potential $\mathbf{P}(p)$, which is intended to measure the attraction at an image point p to the target surface we seek to segment. Note that the image point p is not equivalent to the image voxel, because it is sampled along each trajectory in continuous space with desirable sampling distance for the specific task structure, i.e., the potential $\mathbf{P}(p)$ is not defined on the voxel grid. The potential $\mathbf{P}$ incorporates the prior knowledge of the target surface regarding the intensity $\mathbf{P_I}$, the smoothness $\mathbf{P_S}$ and the geometry $\mathbf{P_G}$. The attraction map is dynamic because the newly resolved vertices are incorporated to update the smoothness term $\mathbf{P_S}$ and geometry term $\mathbf{P_G}$ of the potential, thereby sequentially imposing more appropriate and complete constraints on the surface resolution.

An example definition of the attraction potential $\mathbf{P}(p)$ at image point p is given as follows:

$$\mathbf{P}(p) = \alpha\ \mathbf{P_I}(p) + \beta\ \mathbf{P_S}(p, \boldsymbol{S}) + \gamma\ \mathbf{P_G}(p, \boldsymbol{S}), \quad \forall\ p \in r \in \boldsymbol{U} \tag{3.8}$$

$$\alpha + \beta + \gamma = 1 \tag{3.9}$$

$$0 \leq \alpha, \beta, \gamma \leq 1 \tag{3.10}$$

Where the potential is considered as a weighted sum of the intensity term $\mathbf{P_I}$, the smoothness term $\mathbf{P_S}$ and the geometry term $\mathbf{P_G}$ with weights α, β and γ respectively; $\boldsymbol{U}$ is the set of unresolved trajectories and $\boldsymbol{S}$ is the set of resolved vertices. Note that the potential $\mathbf{P}(p)$ is only defined for image point p located along an unresolved trajectory r, because for resolved

trajectories, the associated surface vertices have already been determined, thus the potential is not needed.

The intensity term $\mathbf{P_I}(p)$ is designed to have a large value if the image point intensity $I(p)$ is close to the expected intensity of the target surface. Note that $I(p)$ is obtained based on trilinear interpolation of the input intensity image, because p may be located between image voxels. If we assume that the surface intensity is generally higher than that of the interior and exterior regions, an example definition of $\mathbf{P_I}(p)$ is given as follows:

$$\mathbf{P_I}(p) = e^{\frac{-\max(0,\ \tau - I(p))^2}{2\sigma_1^2}} \tag{3.11}$$

Where τ is the intensity threshold; σ_1 is the parameter used to control the potential penalty for points with low intensities. However, if the target surface has generally lower intensity than the surrounding regions, (3.11) needs to be modified as:

$$\mathbf{P_I}(p) = e^{\frac{-\max(0,\ I(p) - \tau)^2}{2\sigma_1^2}} \tag{3.12}$$

The smoothness term $\mathbf{P_S}(p, \boldsymbol{S})$ is defined based on the assumption that target surface is smooth enough in the local neighborhood, suggesting that surface vertices tend to be close to each other. An example definition is given as:

$$\mathbf{P_S}(p, \boldsymbol{S}) = e^{\frac{-\mathrm{dist}(p,\ \boldsymbol{S})^2}{2\sigma_2^2}} \tag{3.13}$$

Where $\mathrm{dist}(p, \boldsymbol{S})$ is the Euclidean distance between p and the nearest resolved vertices in $\boldsymbol{S}$; σ_2 is the parameter used to control the potential penalty for points far away from resolved surface vertices.

The geometry term $\mathbf{P_G}(p, \boldsymbol{S})$ incorporates the prior knowledge that the target surface is approximately convex by penalizing points with significantly shorter distance to the centroid $\boldsymbol{C}$ than resolved neighboring vertices. An example geometry term is defined as follows:

$$\mathbf{P_G}(p, \boldsymbol{S}) = \min\left(\frac{dist(p,\ \boldsymbol{C})}{\max\limits_{p' \in \boldsymbol{N}(p),\, p' \in \mathbf{S}} dist(p',\ \mathbf{C})}, 1 \right) \tag{3.14}$$

Where dist(p, $\boldsymbol{C}$) is the Euclidean distance between p and the centroid $\boldsymbol{C}$; $\boldsymbol{N}(p)$ is the set of neighboring surface vertices of p as used in (3.6).

The sequential surface resolution, as summarized in Algorithm 2, resolves one vertex v from $\boldsymbol{F}$ per step by determining the location p with maximal potential $\mathbf{P}(p)$ along the respective trajectory $r(v)$. The order of resolution depends on the amount of local information regarding the target surface. A region with more resolved neighboring vertices provide more evidence of the target surface, because both the smoothness term $\mathbf{P_S}$ and the geometry term $\mathbf{P_G}$, defined in (3.13) and (3.14) respectively, reply on the resolved neighboring vertices. Therefore, the surface resolution is performed in a decreasing order of GNI as defined in (3.6).

Algorithm 2 Sequential surface resolution

while $|\boldsymbol{F}| > 0$ **do**
  maxGNI $\leftarrow \max\limits_{v \in \boldsymbol{F}}$ GNI(v)

    for each $v \in \boldsymbol{F}$ and GNI(v) = maxGNI **do**

      $v \leftarrow \arg\max\limits_{p} \mathbf{P}(p), \quad p \in r(v)$

      remove v from $\boldsymbol{F}$, add v to $\boldsymbol{S}$
    end for

  update GNI(v) for each vertex $v \in \boldsymbol{F}$
end while

3.2.1.2 ROI determination for mean CT attenuation measurement

Two types of 3D volumetric region of interest (ROI) within each segmented vertebral body, corresponding to 1) trabecular bone tissue only (ROI_T) and 2) both trabecular and cortical bone tissues (ROI_C), as illustrated in Figure 3.19 (a-b), have been explored. Previous studies [73, 74, 75, 76, 77, 78, 80, 82, 83, 84] [85, 77, 86, 88, 89] on BMD assessment from CT mostly focus on a trabecular ROI, since trabecular bone tissue is metabolically more active and thus the main determinant of compressive strength in the vertebrae compared to cortical bone tissue [72]. Cortical BMD is included in the measurement of ROI_C, which is expected to better correlate with the reference standard BMD_{DXA}, because BMD_{DXA} is a compound measurement of both cortical and trabecular BMD.

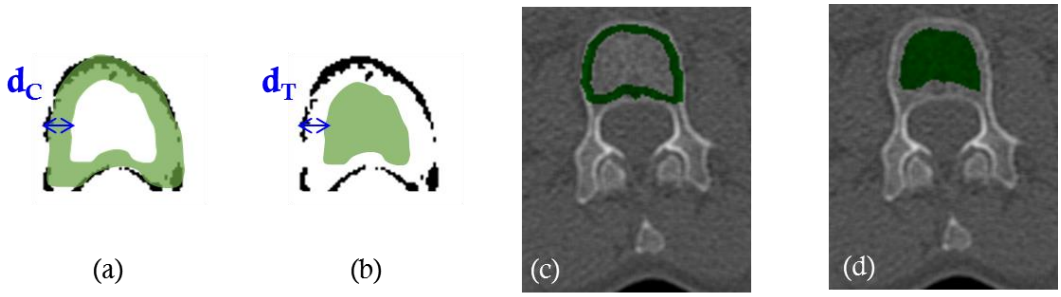

Figure 3.19. 3D volumetric regions of interest within segmented vertebral body. (a) ROI_C (green) consists of both trabecular bone tissue and cortical bone. (b) ROI_T (green) consists of trabecular bone tissue only. tissue. (c) ROI_C (d_C = 3mm) and (d) ROI_T (d_T = 4 mm) shown in green on top of an axial slice of LDCT.

ROI_C is defined as the outer shell of the vertebral body that is within distance d_C to the segmented cortical surface as illustrated in Figure 3.19 (a). ROI_T is determined by excluding outer shell of the vertebral body that is within distance d_T to the cortical surface as illustrated in Figure 3.19 (b). The cortical thickness of human vertebral body is often less than 0.4 mm [91], whereas the spatial resolution of LDCT are typically larger than 0.5 mm, thereby

causing severe partial volume artifacts at the cortical surface. In order to avoid the distortion of measurement at the surface of the vertebral body, pixels within axial distance d_A or vertical distance d_V to the segmented cortical surface are excluded from the ROI. In addition, it is infeasible to extract an ROI consisting of only cortical bone tissue from LDCT, and all the presented ROI_C in this paper are in fact a mixture of both cortical and trabecular bone tissue. Examples of ROI_C and ROI_T are shown in Figure 3.19 (c-d).

3.2.2 Experiments

The segmentation of vertebral bodies by the PSR algorithm was evaluated both qualitatively and quantitatively. A reference level set approach was also applied to the same set of vertebral bodies and compared to the PSR algorithm quantitatively in terms of segmentation accuracy, computation complexity and runtime. The presented BMD_{CT} assessment framework was validated using the BMD_{DXA} as reference standard.

3.2.2.1 Evaluation of the vertebral body segmentation

The PSR algorithm was applied to segment the cortical surfaces of 460 thoracic vertebral bodies from the VIA-ELCAP public database, which consists of 50 low-dose chest CT scans taken for the lung cancer screening purpose. For the evaluation of the algorithm, 4 cases from the database were excluded, due to severe image artifacts caused by medical implants; 10 median vertebral bodies per CT image were segmented and evaluated, which constituted 460 vertebral bodies in total. A single GE CT scanner (GE LightSpeed Ultra, helical mode) was used with tube voltage of 120 KV and slice thickness of 1.25 mm. The in-plane resolution ranges from 0.51mm to 0.82mm. The scans were preprocessed by a 3 x 3 x 1 mean filter and clipped by the bounding box of the input high confidence boundary image $\boldsymbol{B}$.

The surface was first initialized using Algorithm 1 as described in 3.2.1.1 with r_1 = 0.45 and r_2 = 0.55, where the trajectories were generated by subdividing each face of an icosahedron into 64 smaller equilateral triangular elements, corresponding to 642 vertices of the triangular mesh representation. The progressive surface resolution was then performed by following Algorithm 2 with the attraction potential as defined in (3.8) – (3.14) with sampling distance of 0.5 mm, τ = 230 HU, σ_1= 50, σ_2= 8, α = 0.3, β = 0.4 and γ = 0.3. These parameters were selected using an additional training set of 10 low-dose CT scans.

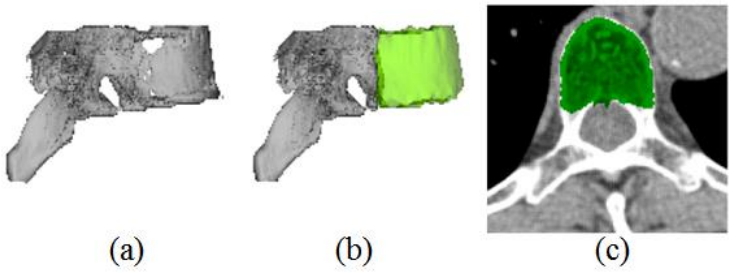

Figure 3.20. The segmentation results were visually inspected in both (a, b) sagittal (the posterior spinal processes are also shown here as reference) and (c) axial view. The 3D visualizations of (a) the input high confidence boundary image ***B*** and (b) the segmented surface (in green) are shown in the sagittal view. (c) The input CT overlaid with the interior region enclosed by the segmented surface (in transparent green) shown in the axial view.

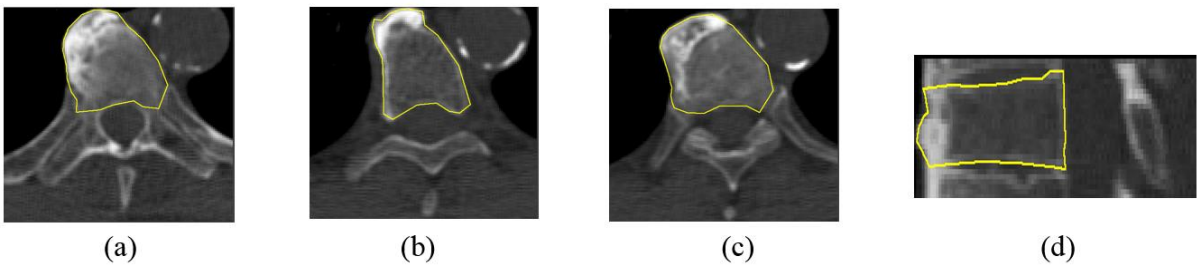

Figure 3.21 Manual annotations of the surface (yellow) used for quantitative evaluations: (a) upper axial slice, (b) median axial slice, (c) lower axial slice, and (d) median sagittal slice.

The segmentation results were first visually inspected in both axial and sagittal views as shown in Figure 3.20. The input high confidence boundary image ***B*** and the resulting

segmentation of the vertebral body surface were compared using their 3D visualizations in sagittal view as shown in Figure 3.20 (a, b). For the axial view, the input CT scan was overlaid with the segmented interior region (i.e., region enclosed by the segmented surface shown in transparent green) at the median axial level as shown in Figure 3.20 (c). The segmentation was considered acceptable if the green regions generally overlapped the vertebral body with no obvious extra structures or missing parts, and there was no hole on the segmented surface.

In addition to the qualitative validation, the segmented surfaces of 46 vertebral bodies (the median vertebra from each CT scan) were also evaluated quantitatively. Manual annotations of the cortical surfaces on the median sagittal slice and three axial slices (upper, median and lower axial slices) were used as the ground truth as shown in Figure 3.21. The upper (lower) axial slice was selected by locating the most superior (inferior) axial slice where more than 80% of the vertebral body cross section appeared. The agreement between the ground truth (***G***, the volume enclosed by the annotated surface) and the segmentation results (***S*,** the volume enclosed by the segmented surface) was measured by the Dice coefficient (DC), which is defined as follows:

$$\mathrm{DC} = \frac{2\,|\boldsymbol{G} \cap \boldsymbol{S}|}{|\boldsymbol{G}|+|\boldsymbol{S}|} \tag{3.15}$$

To compare the PSR algorithm with other alternatives, the 46 vertebral bodies used for quantitative evaluation of the PSR algorithm were also segmented using the level set approach, which was similar to the methods employed by Tan et al. [92]. The implementation was based on the ITK level set segmentation module [93] following the geodesic active contour model proposed by Caselles et al. [94].

For a fair comparison, the same initial seed point and input gray level chest CT were used for both the PSR and level set method. The centroid **C** generated in section 3.2.1.1 was used as the initial seed point for the level set algorithm. The dividing planes described in section 3.2.1.1 for separating individual vertebra and for separating vertebral body from the posterior spinal processes were also used to select the volume of interest in the input CT scans. In addition to the seed point location, 14 parameters are required for the level set algorithm. These parameters were selected using an additional training set of 10 low-dose CT scans. The same quantitative evaluation method as described above was applied.

Table 3.2 Pearson correlation coefficient between BMD_{DXA} and BMD_{CT} measurements at each vertebra level (T1 to L2) and for each ROI (1 to 8). d_A =0.5 mm and d_V =5 mm were used for all eight ROIs. The number of available vertebrae for BMD analysis are specified by n= at each vertebral level. The last three rows correspond to the average BMD_{CT} measurements. All correlation coefficients are statistically significant (p-value < 0.001).

ROI	Trabecular				Cortical & Trabecular			
	1	2	3	4	5	6	7	8
Parameters (mm)	d_T =1.0	d_T =2.0	d_T =3.0	d_T =4.0	d_C =1.0	d_C =3.0	d_C =5.0	d_C =7.0
T1 (n=58)	0.620	0.606	0.569	0.556	0.585	0.637	0.621	0.614
T2 (n=76)	0.740	0.735	0.721	0.718	0.683	0.740	0.741	0.736
T3 (n=76)	0.745	0.725	0.702	0.686	0.732	0.768	0.760	0.757
T4 (n=76)	0.755	0.734	0.713	0.689	0.782	0.800	0.788	0.780
T5 (n=76)	0.750	0.726	0.689	0.649	0.774	0.813	0.796	0.781
T6 (n=76)	0.741	0.737	0.703	0.685	0.733	0.779	0.780	0.770
T7 (n=76)	0.762	0.755	0.735	0.715	0.688	0.780	0.783	0.776
T8 (n=76)	0.776	0.775	0.755	0.723	0.728	0.804	0.801	0.793
T9 (n=76)	0.775	0.759	0.739	0.709	0.764	0.833	0.829	0.820
T10 (n=76)	**0.807**	0.787	0.757	0.718	0.782	**0.850**	**0.850**	0.844
T11 (n=76)	0.792	0.775	0.750	0.729	0.785	0.849	0.848	0.838
T12 (n=72)	0.721	0.706	0.669	0.643	0.779	0.812	0.804	0.789
L1 (n=60)	0.695	0.657	0.620	0.594	0.687	0.778	0.764	0.747
L2 (n=38)	0.605	0.621	0.568	0.546	0.623	0.678	0.651	0.633
Ave T3-T8 (n=76)	0.789	0.778	0.751	0.726	0.773	0.823	0.818	0.810
Ave T9-T11 (n=76)	**0.809**	0.793	0.772	0.744	0.794	**0.857**	0.855	0.848
Ave L1-L2 (n=38)	0.720	0.719	0.672	0.649	0.759	0.788	0.762	0.746

3.2.2.2 Evaluation of the BMD_{CT} assessment framework

The presented BMD_{CT} assessment framework was validated using a dataset consisting of 76 pairs of DXA and LDCT scans of the same subject. Most of the subjects underwent the DXA and LDCT within 2 months (61% within 2 months; 9% between 2 months and a year; 16% between a year and 2 years; and 14% > 2 years). The average aBMD (in g/cm^2) of L1 to L4, BMD_{DXA}, was obtained using the manufacturer-supplied software and GE Lunar iDXA DXA scanner in the postero-anterior view by a board-certified radiologist. All LDCT scans were acquired on Siemens CT. Various scanner models and reconstruction kernels were used.

Vertebral bodies were segmented and anatomical labeled from all 76 LDCT scans using the presented fully automated framework. Eight different ROIs, including four ROI_T and four ROI_C, as defined by the parameters (d_T, d_C, d_A, and d_V) specified in Table 3.2, were constructed for each segmented vertebral body. The mean CT attenuation in HU was measured within each ROI and considered as the corresponding BMD_{CT} at the respective vertebral level. The average BMD_{CT} of upper thoracic vertebrae T3-T8, lower thoracic vertebrae T9-T11, and lumber vertebrae L1-L2 were also calculated. The Pearson correlation coefficient was computed between each BMD_{CT} measurement and the reference BMD_{DXA}. P-values < 0.001 were accepted as statistically significant.

3.2.3 Results

3.2.3.1 Evaluation of the vertebral body segmentation

For the qualitative evaluation, the PSR algorithm was able to achieve acceptable surface segmentation of 99.35% (457 out of 460) vertebral bodies as shown in Figure 3.22 (a-

c), whereas only 0.65% (3 out of 460) vertebral bodies were segmented incorrectly as shown in Figure 3.22 (d-f).

The quantitative evaluation results of 46 vertebral bodies are summarized in Table 3.3. The automated segmentation is shown to be consistent with the manual annotated ground truth with the mean Dice coefficient (DC) larger than 0.93 on both the axial (including upper, median and lower) and the median sagittal slices. Four examples of the segmentation with different DC values are shown in Figure 3.23.

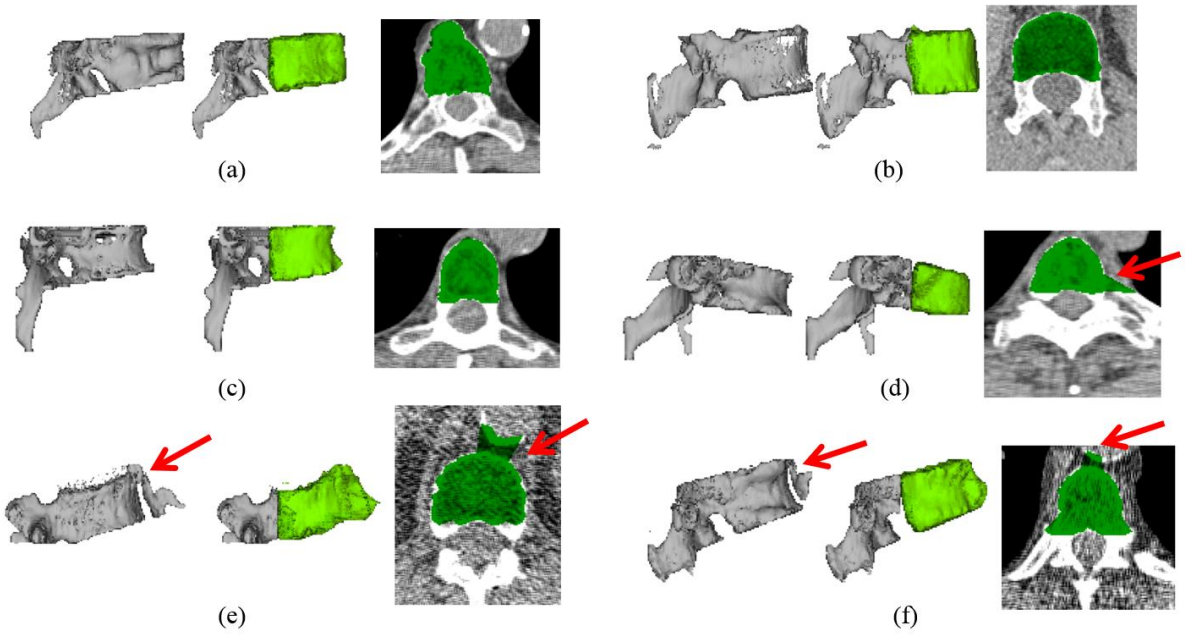

Figure 3.22 Three examples of acceptable segmentation (a-c) and three unacceptable cases (d-f). The arrows indicate the structures incorrectly segmented.

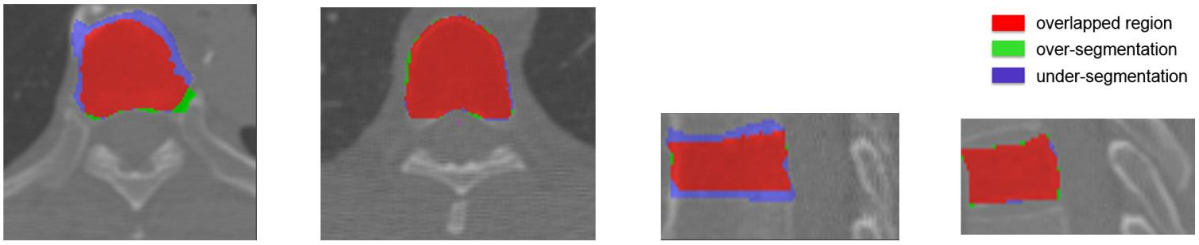

Figure 3.23 Four examples of the segmentation results. (a) DC_L = 0.870, lower axial slice. (b) DC_M = 0.975, median axial slice. (c) DC_S = 0.834, median sagittal slice. (d) DC_S = 0.980, median sagittal slice.

The quantitative comparison results of the segmentation obtained by the PSR algorithm and the level set approach for 46 vertebral bodies are summarized in Table 3.4. The PSR algorithm was able to segment a vertebral body with a mean DC of 0.939 in 0.0443 seconds on average; while the level set approach achieved the segmentation with a mean DC of 0.875 in 11.66 seconds on average. The paired t–test results for the DC measurements in Table 3.4 indicated that the PSR has a statistically significant improved segmentation performance compared to the level set method ($p < 0.001$) for each of the four DC measures.

Table 3.3 The quantitative evaluations results for 46 vertebral bodies. Dice coefficient (DC) for the upper axial slice (DC_U), the median axial slice (DC_M), the lower axial slice (DC_L) and the median sagittal slice (DC_S) are reported below.

	DC_U	DC_M	DC_L	DC_S	Mean
Mean	0.933	0.946	0.936	0.942	0.939
Max	0.963	0.975	0.962	0.980	0.957
Min	0.872	0.906	0.870	0.834	0.906
Standard Deviation	0.022	0.013	0.017	0.022	0.011

Table 3.4 The quantitative comparison results. The average Dice coefficient (DC) for the upper axial slice (DC_U), the median axial slice (DC_M), the lower axial slice (DC_L), the median sagittal slice (DC_S), the mean DC, and the average execution time (per vertebra) are reported below.

	DC_U	DC_M	DC_L	DC_S	Mean DC	Execution time
PSR	0.933	0.946	0.936	0.942	0.939	0.0443 s
Level set	0.873	0.864	0.885	0.876	0.875	11.66 s

3.2.3.2 Evaluation of the BMD_{CT} assessment framework

The number of available vertebra bodies varies from scan to scan due to the difference in scan vertical coverage. As shown in the first column in Table 3.2, T2 – T11 are available for BMD analysis on all 76 scans; T12 is available on 94.7% of the scans; T1 is available on 76.3% of the scans; L1 is available on 78.9% of the scans; L2 is available on 50% of the scans.

The Pearson correlation coefficient between BMD_{DXA} and BMD_{CT} measurements at each vertebra level (T1 – L2) and for each ROI (1-8) are summarized in Table 3.2. The last three rows correspond to the average BMD_{CT} measurements. All correlation coefficients are statistically significant (p-value < 0.001). The Pearson correlation coefficients between BMD_{CT} and BMD_{DXA} at different vertebra levels using the best ROI_T (ROI 1) and best ROI_C (ROI 6) are shown in Figure 3.24 (a) and (b) respectively. The Pearson correlation coefficients between BMD_{DXA} and the average BMD_{CT} of T3-T8, T9-T12, and L1-L2 respectively for all eight ROIs are shown in Figure 3.25.

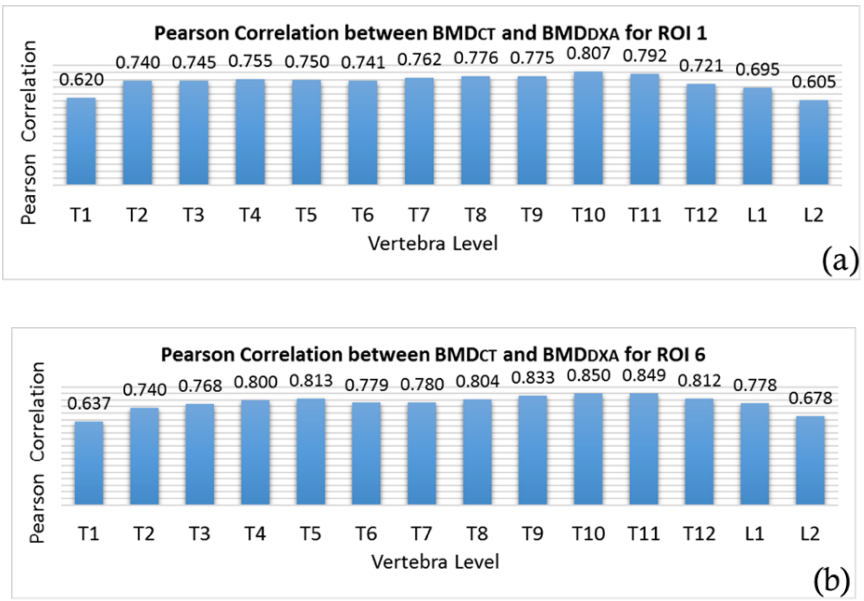

Figure 3.24. Pearson correlation coefficient between BMD_{CT} and BMD_{DXA} at different vertebra levels for (a) trabecular ROI 1 and (b) cortical ROI 6.

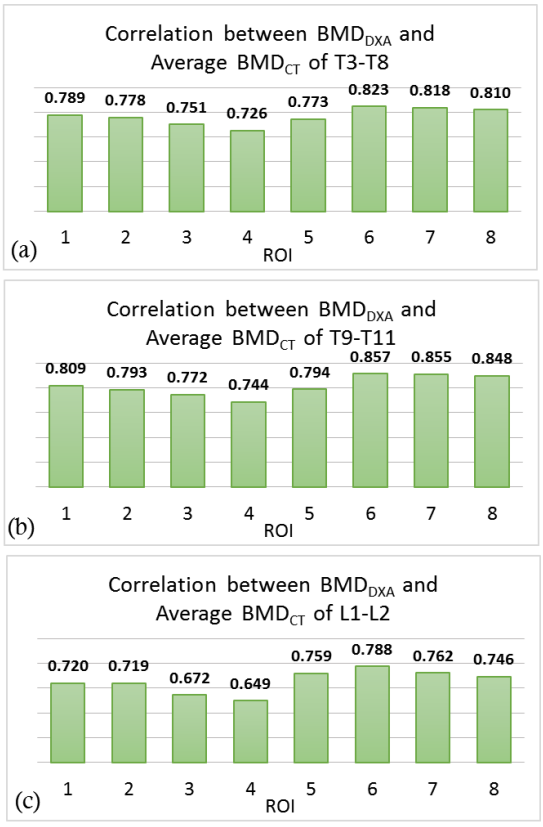

Figure 3.25. Pearson correlation coefficient between BMD_{DXA} and average BMD_{CT} of (a) T3-T8, (b) T9-T11, and (c) L1-L2 for eight ROIs.

3.2.4 Discussion

Both qualitative and quantitative validations of the vertebral body surface segmentation demonstrate the robustness and promising performance of the PSR algorithm. The primary reason caused the 0.65% (3 out of 460) incorrect surface segmentation was external calcifications, i.e., left rib in Figure 3.22 (d) and aorta calcifications in Figure 3.22 (e, f), which are located very close to the vertebral body and are incorrectly included in the input high confidence boundary $\boldsymbol{B}$. For the attraction potential and the surface initialization rules, there is an underlying assumption that $\boldsymbol{B}$ provides no significant calcifications other than the target vertebra itself. Therefore, the algorithm is capable of dealing with the spurious boundaries in the interior region and the holes on the cortical surfaces; however, it is not designed to perform robustly when $\boldsymbol{B}$ contains significant calcifications outside the vertebra, which is the case as shown in Figure 3.22 (d-f). Fortunately, the automated method we employed to obtain $\boldsymbol{B}$ provides high confidence boundaries satisfying the above assumption in almost all cases (99.35%, 457 out of 460).

Additional modifications to the presented PSR algorithm are required to further refine the three unacceptable cases shown in Figure 3.22 (d-f). This issue may be addressed either by additional preprocessing to remove these structures prior to algorithm execution or by modifying the vertebra specific modifications to the PSR algorithm. The latter can be achieved by new distance exclusion rules for large distances or by adding additional factors to the attraction potential formulation. For the future, we will consider a larger dataset to collect sufficient cases with similar problem conditions, and explore the algorithm refinements mentioned above.

The PSR algorithm is able to provide a robust outcome with improved computation complexity (no iterations required), as compared to the deformable models that have been widely applied to segment a closed surface. The surface resolution process of the PSR is sequential, and the location of each vertex is optimized along a single radial trajectory. Therefore, the computation complexity and runtime are linearly constrained by the total number of vertices on the surface mesh (642 vertices employed in our implementation). For our evaluation of the level set method shown in Table 3.4, this iterative method was more than 250 times slower than the proposed PSR algorithm. The outcomes for the level set method were inferior to the PSR method (average Dice coefficient of 0.875 compared to 0.939).

As pointed out by Tan et al. [92], there are many parameters to be set for the level set method, which are difficult to tune. For algorithm optimization we individually varied all parameters through several iterations. The possibility of some improved performance with fewer algorithm iterations may be achievable. Also, incorporating multi-scale methods as suggested in the preliminary study [92] may address some of the leakage issues that were observed with the level set method. However, we observed a significant improvement with respect to speed and accuracy of our proposed method that is unlikely to be accounted for by parameter setting optimization.

BMD_{CT} measurements obtained using the presented fully automated framework have been demonstrated to have strong statistically significant correlation, as shown in Table 3.2, with BMD_{DXA}, which is currently the most widely used and gold standard technique for the diagnosis of osteoporosis and fracture risk estimation [72]. Among the four ROI_T, the greatest correlation of 0.809 is obtained using the average BMD_{CT} of the vertebra level T9-T11, which

is comparable to the best performance reported by Lee et al. in [85] (r=0.880 at L3 from abdominal routine CT scan) and by Kim et al. in [86] (r = 0.726 for average of T4, T7, T10 and L1 from LDCT). The reason that our performance is worse than that reported by Lee et al. in [85] may be due to the much higher level of image noise in LDCT compared to routine abdominal CT in general.

BMD_{CT} measurements including cortical bone tissue (i.e., using ROI_C) have been demonstrated to have better correlation compared to just considering trabecular bone tissue (i.e., using ROI_T), as shown in Table 3.2, Figure 3.24 and 3.25. Among the four ROI_C, the greatest correlation of 0.857 is obtained between the average BMD_{CT} of T9 –T11 and BMD_{DXA}. All four ROI_C are in fact a mixture of both cortical and trabecular bone tissue, which explains its greater correlation with BMD_{DXA} that is also a compound measurement of both cortical and trabecular bone tissue. Similar results from routine abdominal CT have been presented by Tay et al [79]. Although better accordance with BMD_{DXA} is obtained by including cortical tissue (mainly due to the inherent limitation of DXA scan), trabecular ROIs should be preferred for the BMD assessment from CT, since it is a more sensitive determinant of compressive strength in the vertebrae [72, 70].

The BMD_{CT} measurements at lower thoracic vertebral levels (T9-T11) have, in general, better correlation with BMD_{DXA} compared to BMD_{CT} measurements at upper thoracic vertebral levels (T3-T8) and lumber levels (L1-L2), as shown in Table 3.2, Figure 3.24 and 3.25. For instance, using ROI 6, the correlation between BMD_{DXA} and average BMD_{CT} of T9-T12 is 0.857, which is better than 0.823 using average BMD_{CT} of T3-T8 and 0.788 using average BMD_{CT} of L1-L2. Since BMD_{DXA} is measured from L1-L4, its correlations with BMD_{CT} are expected to be higher at vertebral levels T9-T12 compared to

T3-T8, as observed by Miyabara et al [75] and Hayashi et al [78]. On the other hand, the lower correlation at lumbar vertebral levels L1-L2 may be caused by the much higher level of image noise that is often observed at axial levels inferior to lungs on the LDCT scans.

With the recent large-scale implementation of annual lung cancer screening in the US using LDCT [10], great potential emerges for the concurrent extraction of quantitative image biomarkers from the chest regions other than lungs, such as heart [95], spine, ribs and breasts [45], which are also covered in LDCT scans. Kim et al. have demonstrated the feasibility of assessing BMD from LDCT using manually delineated ROI [86]. The results presented in this paper shows the feasibility of fully automated BMD assessment as well as the potential of osteoporosis diagnosis and vertebral compression fracture detection from LDCT acquired in the annual lung cancer screening.

3.2.5 Conclusion

A fully automated system is presented in this study for the BMD assessment based on image attenuation (HU) from the LDCT acquired during annual lung cancer screening. Average BMD of L1-L4 measured from DXA are used as the reference standard for the validation of BMD_{CT}. Statistically significant (p-value < 0.001) strong correlation can be obtained between BMD_{DXA} and BMD_{CT} at all vertebral levels (T1 – L2). The highest Pearson correlation of 0.857 is achieved between BMD_{DXA} and the average BMD_{CT} of T9-T11 by using a 3D ROI considering both trabecular and cortical bone tissue. The encouraging results demonstrate the feasibility of fully automated quantitative BMD assessment and the potential of opportunistic osteoporosis screening with concurrent lung cancer screening using LDCT.

CHAPTER 4

FULLY AUTOMATED AIRWAY LABELING AND QUANTITATIVE BIOMARKER MEASUREMENTS

Chronic obstructive pulmonary disease (COPD), a heterogeneous disease associated with varying degrees of emphysema and small airways disease [21], is expected to be the 3rd leading cause of death by 2020 [22]. Several studies have established that airway dimensions (such as lumen diameter and wall thickness) are related to the airflow obstruction and peripheral airway inflammation in patients with COPD [96, 97, 98].

Computed tomography (CT) is widely performed for patients with COPD [99]. The airway dimensions measured from CT chest scans have been demonstrated repeatedly to be correlated with measures of airflow obstruction [100, 101, 102, 103, 104, 105, 106, 107, 108]. CT scanning allows for quantitative assessment of airway dimensions in vivo [109], which might facilitate more accurate diagnosis and treatment, COPD phenotypic dissection [109, 110, 111], as well as the evaluation of novel therapies [21]. Airways smaller than 2-mm in internal diameter are the main sites of airflow obstruction in patients with COPD [97, 112, 113]; however, they are located from the 4^{th} to the 14^{th} generation of airway branching in the lung [97], and thereby are difficult to measure from CT scans due to limited scan resolution and high levels of image noise especially when low radiation dose is used. Fortunately, it has been demonstrated that CT measurements of larger airways (such as 3^{rd} - 6^{th} generations), which can be assessed more precisely from CT, also correlate well with the severity of COPD, and may be used as a useful predicator of small airway pathology [101, 96, 100, 114, 103, 111, 104, 106, 115]. This is because that the same pathophysiologic process may occur at all

levels of the airway tree [96], although the increased airway wall thickness and the narrowing in the larger airways may have no profound functional consequences [115].

Most research on the quantification of airway dimensions by CT scans to date still involves manual editing during the process of airway segmentation, anatomical labeling as well as the measurement of airway dimensions [104, 98, 100, 101, 114, 102, 103, 105, 115, 116]. Various assessment parameters have been proposed in these semi-automated studies, including 2D cross sectional measurements such as wall thickness, lumen area and wall area percentage, 3D volume based measurements such as wall volume and lumen volume [105], branching generation number [105], and intensity based parameter such as peak wall attenuation [101]. Wall thickness, lumen area and wall area percentage are the most commonly adopted measurements, which were initially acquired from axial CT image slices [98, 100, 101, 115] and introduced inaccuracy due to the fact that the airway long axis can be oblique to the imaging plane. Recent studies [114, 102, 103, 104, 105, 116] circumvented this issue by adopting 3D approaches to first reconstruct airway cross sections orthogonally along the airway long axis before performing the measurement. All of the aforementioned semi-automated approaches require manual intervention, thereby can be time-consuming, prone to considerable inter- and intra-observer variability and dependent on the display parameters of the CT image (i.e., window settings) [110, 117].

A number of fully automated studies [107, 106, 108, 118] have also been conducted to assess airway dimensions from CT images, generally including stages of airway segmentation, centerline extraction and dimension measurements. Airway dimensions and characteristics vary significantly for different regions of the lung and for different generations [118], therefore performing reproducible and comparable airway measurements requires the

ability to locate the corresponding positions to measure in different scans, which is challenging in fully automated studies. Two main schemes have been proposed to find corresponding airway locations to perform the airway assessment and enable the inter-patient comparison in a cross-sectional study and intra-patient evaluation in a longitudinal study. One approach is averaging all detectable airway dimensions at a predefined internal lumen perimeter (or equivalently a predefined inner lumen diameter) [106, 115], which has the limitation that the measured changes may be due to modifications either in airway lumen, airway wall, or both [117], and fails to provide regional (or generation specific) assessment. The second approach require automated airway categorizing or labeling [117], so that the dimensions can be assessed either by averaging at a special anatomical generation [108, 107] or at a specific anatomical bronchus [118].

The fully automated assessments [107, 108] usually rely on topological generation that is defined by the topology of the segmented airway tree: starting from trachea that is of topological generation 0, each bronchus of topological generation n, may branch into several descendant bronchi of generation n+1. However, all the non-automated studies discussed before investigated the anatomical generation that is defined by the anatomical functions of airway bronchi as illustrated in Table 4.1, which prevents the direct comparison between most of the automated studies and the clinical studies. Moreover, the studies with the ability of automated anatomical airway labeling to date have been able to measure airway dimensions as far as the 3rd anatomical generation (segmental levels) [118], whereas the semi-automated studies can investigate airway up to 6th anatomical generations and demonstrate that the more peripheral airway generations usually provide a stronger predictor for airflow obstructions [101, 114, 103]. Therefore, a fully automated framework that allows quantification of airway

dimensions for specific anatomical bronchi and for averaging over further anatomical generations is still needed to facilitate the diagnosis and treatment of COPD.

The purpose of this study was to develop a fully automated anatomy directed framework for the analysis of reproducible and comparable quantitative airway biomarkers from low-dose chest CT (LDCT) scans. The airway is first segmented with each bronchus labeled with corresponding anatomical name following the nomenclature defined in [119]. Then the lumen diameter and wall thickness of each bronchus are measured at each bronchus, serving as reproducible and comparable biomarkers that provide valuable information aiding the diagnosis and treatment of COPD.

4.1 Framework for airway anatomical labeling and quantitative biomarker measurements

The fully automated anatomy directed framework consists of three main stages as illustrated in Figure 4.1. First, the airway is segmented with its centerline modeled by a tree structure. Second, the lumen diameter and wall thickness are measured at each of the bronchi (tree branches). Third, each segmented bronchus is labeled with its corresponding anatomical name.

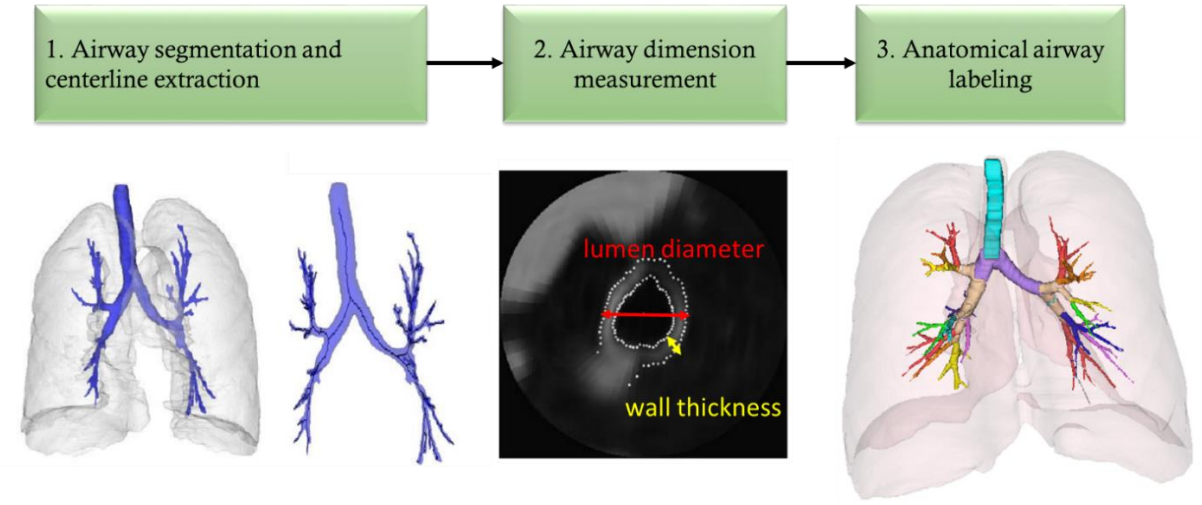

Figure 4.1. Flow chart of the presented framework for the quantitative airway image biomarker analysis.

The airway segmentation and the tree structure formed by the airway centerline is obtained in the first stage of the framework. The airway centerline algorithm was proposed by Lee et al. in [67, 120] and consists of three main steps. Frist, the airway is segmented from the LDCT scan by 3D region growing within a locally-defined envelope. The airway centerline is then obtained by applying 3D thinning to the segmented airway. Finally, a tree structure is used to model the airway centerline with individual airway bronchus identified based on the analysis of the hierarchy of extracted centerline.

The airway dimensions, including the wall thickness and lumen diameter are measured in the second stage of the framework by following the approach proposed by Lee et al. in [67]. The location of the inner wall and the outer wall is determined for an airway bronchus based on the full width at half maximum (FWHM) principle, where the inner wall is defined as the boundary between the airway lumen and wall and the outer wall is defined as the boundary between the airway wall and lung parenchyma. The lumen diameter and the wall

thickness are then derived from the volume and length measurement of each bronchus with the assumption that each bronchus can be approximated by a cylinder.

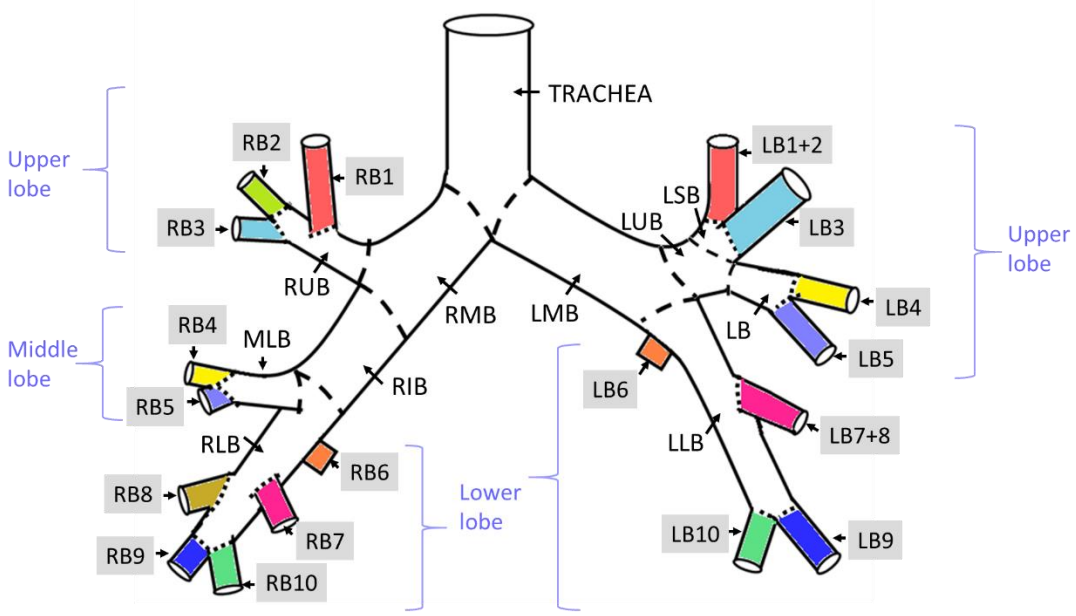

Figure 4.2. The illustration of airway bronchi of anatomical generation 0 – 3 and abbreviations used in this study.

Each individual bronchus is identified and assigned with its anatomical name in the final stage of the framework. The fully automated anatomical labeling of airway bronchi plays an important role in order to enable reproducible and comparable quantitative airway biomarkers analysis. It allows the biomarker measurements to be reported for a specific bronchus or to be averaged over a specific anatomical generation, leading to direct comparison of the biomarker measurements of a patient with that of the healthy subject (reference standard), as well as the direct comparison of the biomarker measurements of the same subject in longitudinal scans, which may aid the diagnosis and therapy monitoring of an extensive variety of airways diseases. In additional, it also allows lobe-based lung heath

analysis, as the anatomically labeled airway bronchi potentially formulate the frame of location reference in the lung regions.

The anatomical labeling of airway bronchi is challenging due to two main reasons. First, there is a large variation of the same airway bronchus across individuals in terms of length, diameter, running direction, locations, etc., either due to normal individual anatomy variations or due to the airway pathology. Second, the airway segmentation errors, such as under-segmentation or over-segmentation, and the inaccurate airway tree formulation from the airway centerline can also increase the difficulty of the airway labeling.

To deal with the anatomy variations and the possible issues resulting from the segmentation and skeletonization step, the airway labeling algorithm employs a hierarchical labeling order based on anatomy variations and a dynamic airway tree model to allow flexibility in the input tree topology. The bronchi with small anatomical variations and prominent features are labeled first. The labeled bronchi are then incorporated into the knowledge base to aid the labeling of subsequent bronchi. Only one feature, the running direction that is relatively unaffected by pathology, is used in the labeling process, except the labeling of RLB, which also used the radius of the bronchi, resulting in more robust labeling in patient with airway diseases. A dynamic airway tree model is adopted by the algorithm, which is constantly adjusted in the labeling process whenever a false bronchus or missing bronchus is identified, to improve the robustness of airway labeling with respect to anatomy variations and the preceding segmentation issues.

A novel approach based on an anatomy directed knowledge base was employed to label each identified airway bronchus with the corresponding anatomical name. The

segmented airway bronchi are labeled in a hierarchical fashion following the topological ordered defined in the airway tree obtained in the previous step. The algorithm is designed to identify 29 anatomical airway bronchi of anatomical generation 0 – 3, as summarized in Table 4.1 and illustrated in Figure 4.2. The labeled include 1 zeroth generation bronchus, 2 first generation bronchi, 8 second generation bronchi, and 18 third generation bronchi. Bronchi of the fourth to ninth generations are labeled with its third-generation ancestor and the anatomical generation number. As a consequence, the algorithm is not designed to differentiate bronchi of the fourth anatomical generation and above with the same third-generation ancestor, which is consistent with convention employed in the literature [106, 107, 108, 118]. The adopted nomenclature follows that defined by Netter in [119].

Table 4.1. Airway bronchi of anatomical generation 0 – 3.

Anatomical generation	Bronchus name	Abbreviation	Lobe	Color
0	Trachea	TRACHEA		
1	Right main bronchus	RMB	Right upper lobe	
1	Left main bronchus	LMB	Light upper lobe	
2	Right upper lobe bronchus	RUB	Right upper lobe	
2	Right intermediate bronchus	RIB	Right upper lobe	
2	Middle lobe bronchus	MLB	Right middle lobe	
2	Right lower lobe bronchus	RLB	Right lower lobe	
2	Left upper lobe bronchus	LUB	Left upper lobe	
2	Left superior division bronchus	LSB	Left upper lobe	
2	Lingular bronchus	LB	Left upper lobe	
2	Left lower lobe bronchus	LLB	Left lower lobe	
3	Right bronchus 1	RB1	Right upper lobe	
3	Right bronchus 2	RB2	Right upper lobe	
3	Right bronchus 3	RB3	Right upper lobe	
3	Right bronchus 4	RB4	Right middle lobe	
3	Right bronchus 5	RB5	Right middle lobe	
3	Right bronchus 6	RB6	Right lower lobe	

3	Right bronchus 7	RB7	Right lower lobe	
3	Right bronchus 8	RB8	Right lower lobe	
3	Right bronchus 9	RB9	Right lower lobe	
3	Right bronchus 10	RB10	Right lower lobe	
3	Left bronchus 1+2	LB1+2	Left upper lobe	
3	Left bronchus 3	LB3	Left upper lobe	
3	Left bronchus 4	LB4	Left upper lobe	
3	Left bronchus 5	LB5	Left upper lobe	
3	Left bronchus 6	LB6	Left lower lobe	
3	Left bronchus 7+8	LB7+8	Left lower lobe	
3	Left bronchus 9	LB9	Left lower lobe	
3	Left bronchus 10	LB10	Left lower lobe	

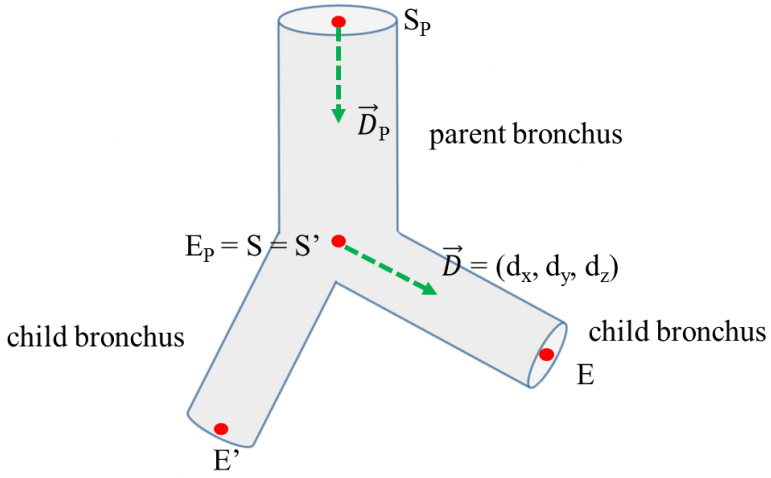

Figure 4.3. Illustration of the parent-children bronchi hierarchy, the start and end of a bronchus and the running direction.

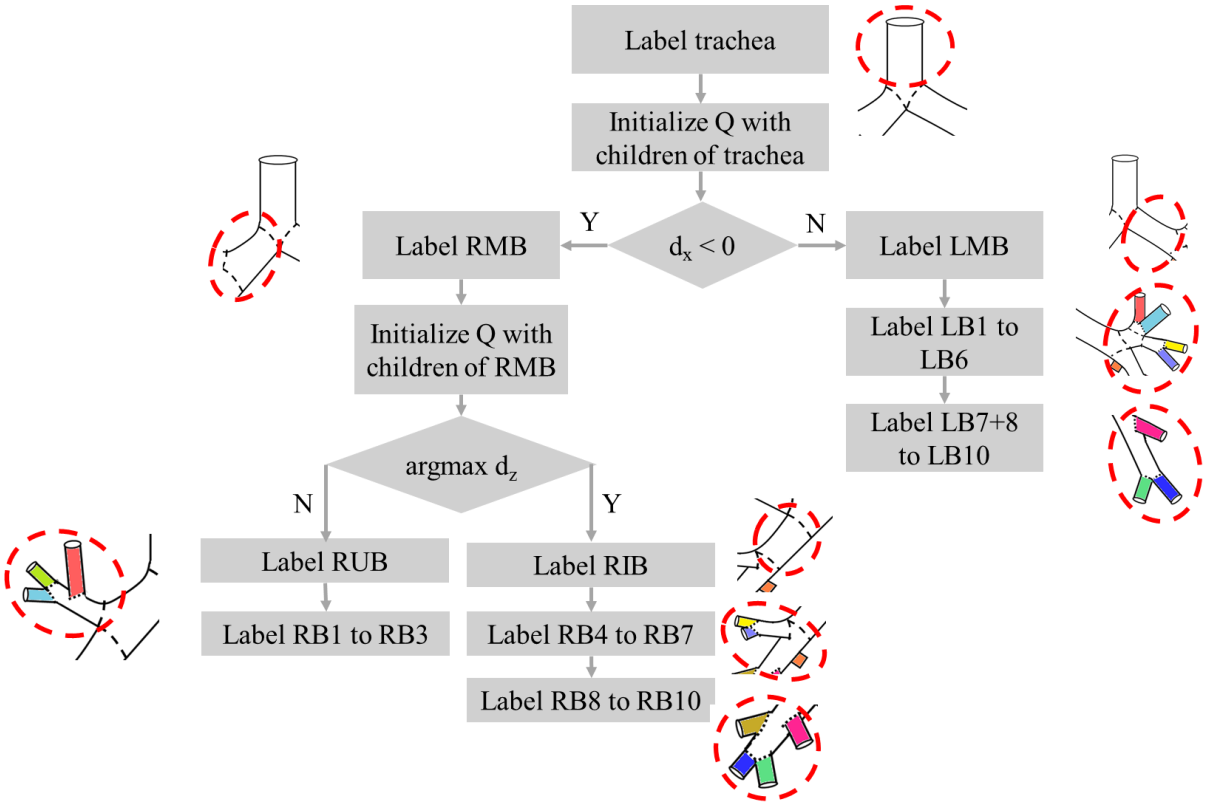

Figure 4.4. Flow chart of the presented algorithm for airway anatomical labeling.

Each segmented airway bronchus is represented by its start coordinates $\vec{S}$, end coordinate $\vec{E}$, which are obtained in the airway skeletonization process, and its radius r that is obtained in the airway dimension measurement process as illustrated in Figure 4.3. Except for the trachea, each bronchus has a unique parent bronchus. Except for the terminal bronchi, each bronchus has at least 2 child bronchi. The parent-child hierarchy can be derived based on the overlapping of start coordinates $\vec{S}$ and end coordinate $\vec{E}$. The absolute running direction of a bronchus is defined as a unit vector $\vec{D}$:

$$\vec{D} = (d_x, d_y, d_z) = (\vec{E} - \vec{S}) / |\vec{E} - \vec{S}| \quad (4.1)$$

The relative running direction of a bronchus is defined as $\Delta\vec{D}$:

$$\Delta\vec{D} = (\Delta d_x, \Delta d_y, \Delta d_z) = \vec{D} - \vec{D}_p \quad (4.2)$$

where $\vec{D}_p$ is the absolute running direction of its parent bronchus.

The labeling process is conducted in a hierarchical fashion defined by the airway tree topology as illustrated in Figure 4.4. A processing queue Q is employed to record the child bronchi of the most recently labeled bronchus to maintain the labeling order. As a result, all candidates in Q have the same parent bronchus with known anatomical label. The candidates are then matched to the anatomy-oriented airway model using only two features: the running direction and the radius of a bronchus. Given the labeled parent bronchus, the child bronchi with small anatomical variations are identified first, and their presence/absence status is taken into considerations during labeling of other siblings to resolve confusion.

The algorithm allows anatomy variations and deals with preceding issues resulting from the segmentation and skeletonization step by allowing flexibility in input tree topology. Each airway bronchus has a presence/absence status that impacts the anatomy model employed to match the subsequent bronchi to be labeled. The algorithm does not require the presence of any bronchus (except trachea); once a bronchus is considered absence, then all its descendants are considered absence. A false bronchus is detected if it satisfies either of the following criteria:

(i) a zeroth to second generation bronchus with length < L_{min};

(ii). a candidate in the processing queue Q, which cannot be matched to any of children of its labeled parent.

The false bronchi can be caused by two main reasons: the subject has extra bronchi that are not identified in the model; the segmentation errors or the tree skeletonization errors. The

false bronchus is ignored in the labeling process. If it satisfies criterion (i), the topological generation number of respective descendants are decreased accordingly.

The flowchart of the algorithm is illustrated in Figure 4.4. The trachea is first labeled as it is the most superior bronchus, which can be trivially identified. The children of trachea are then labeled as RMB and LMB based on d_x. The descendants of RMB are then labeled sequentially as RB1 to RB3 as illustrated in Figure 4.5, as RB4 to RB7 as illustrated in Figure 4.6, and as RB8 to RB10 as illustrated in Figure 4.7. The descendants of LMB are then labeled sequentially as LB1 to LB6 as illustrated in Figure 4.8 and as LB7+8 to LB10 as illustrated in Figure 4.9.

The absence of a bronchus and the different number of segmented child bronchi of a bronchus may lead to the variations of the airway model to be matched to in the subsequent labeling process. There are five major possible variations of the employed airway model:

(1). The algorithm allows variation in the topological order of RUB and RB1, as illustrated in Figure 4.5. If RUB has two children, then one child bronchus should be RB1 and the other should be the extended RUB; On the other hand, if RUB has more than two children, then the extended RUB is considered absent.

(2). The algorithm allows variation in the topological order of MLB, RB6 and RB7, as illustrated in Figure 4.6. Among the children of RIB, one must be RLB, and the others can be MLB, RB6 or RB7. If the processing queue Q is empty and one of MLB, RB6 and RB7 is unlabeled, then Q is reinitialized by children of recently labeled RLB, and recursively apply the current step to label (extended) RLB, MLB, RB6 and RB7.

(3). The algorithm allows variation in the topological order of RLB and RB8, as illustrated in Figure 4.7. After the labeling of RB1 to RB7, if RLB has two children, then one child bronchus should be RB8 and the other should be the extended RLB. On the other hand, if RLB has more than two children, then the extended RLB is considered absent.

(4). The algorithm allows variation in the topological order of LLB, LUB, LB6, LSB, and LB, as illustrated in Figure 4.8. If LMB has two children, then they are considered to be LUB and LLB. On the other hand, if LMB has more than two children, they can be any one of LLB, LUB, LB6, LSB, and LB. The anatomical variation of LLB and LB6 are relatively small, hence they are matched first. If one child of LMB is labeled as LB, then LUB is considered absent. If no child of LMB is labeled as LB6, then LB6, if present, is sought among children of LLB. If LUB is labeled, then LB is sought among it children. LSB may be absent depending on the number of candidates in the processing queue Q after the labeling of LB.

(5). The algorithm allows variation in the topological order of LLB and LB7+8, as illustrated in Figure 4.9. After the labeling of LB1 to LB6, if LLB has two children, then one child bronchus should be LB7+8 and the other should be the extended LLB. On the other hand, if LLB has more than two children, then the extended LLB is considered absent.

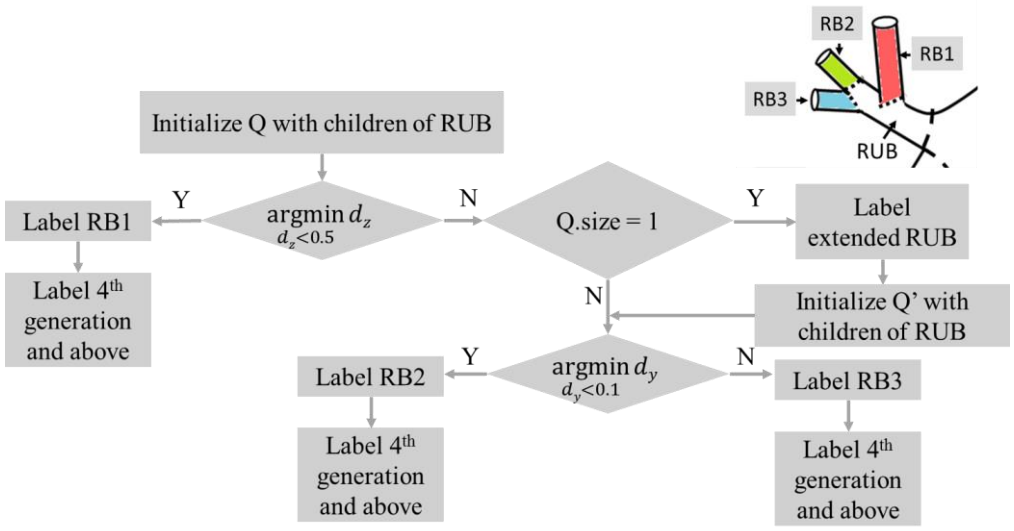

Figure 4.5. Flow chart algorithm of labeling RB1 to RB3.

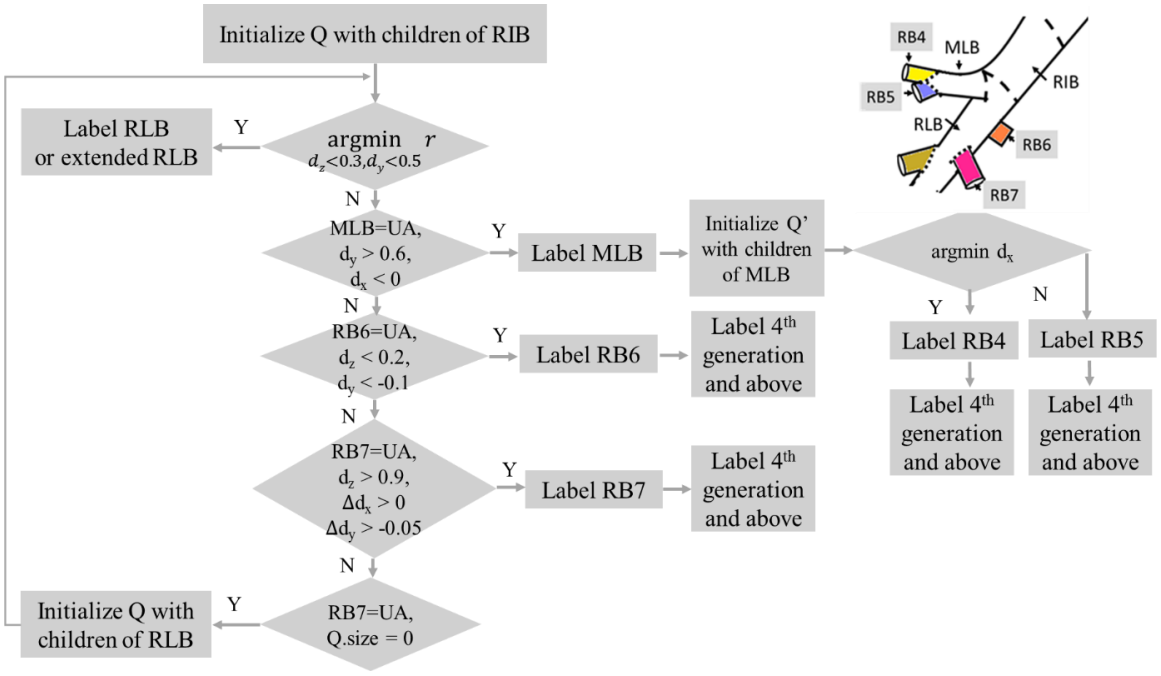

Figure 4.6. Flow chart algorithm of labeling RB4 to RB7. Unassigned label is denoted by UA.

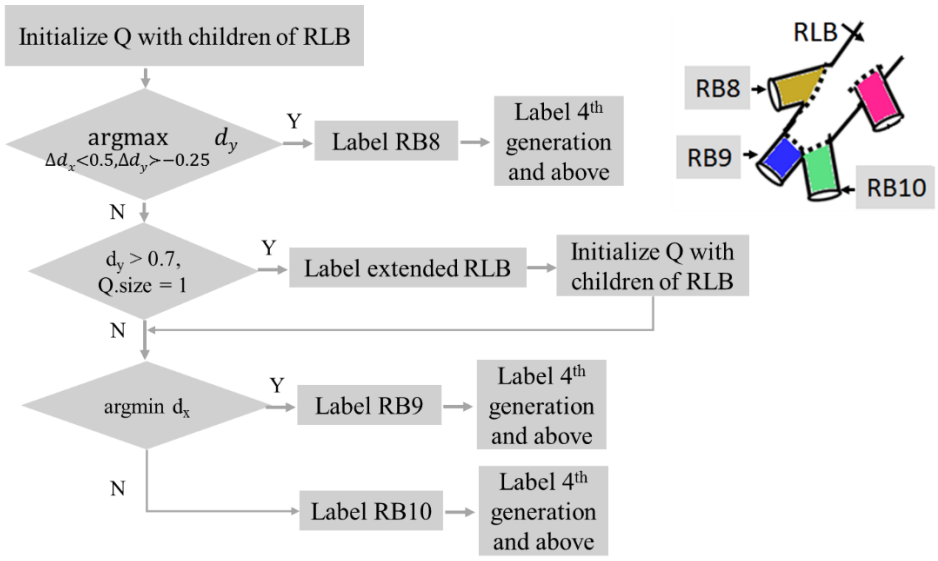

Figure 4.7. Flow chart algorithm of labeling RB8 to RB10.

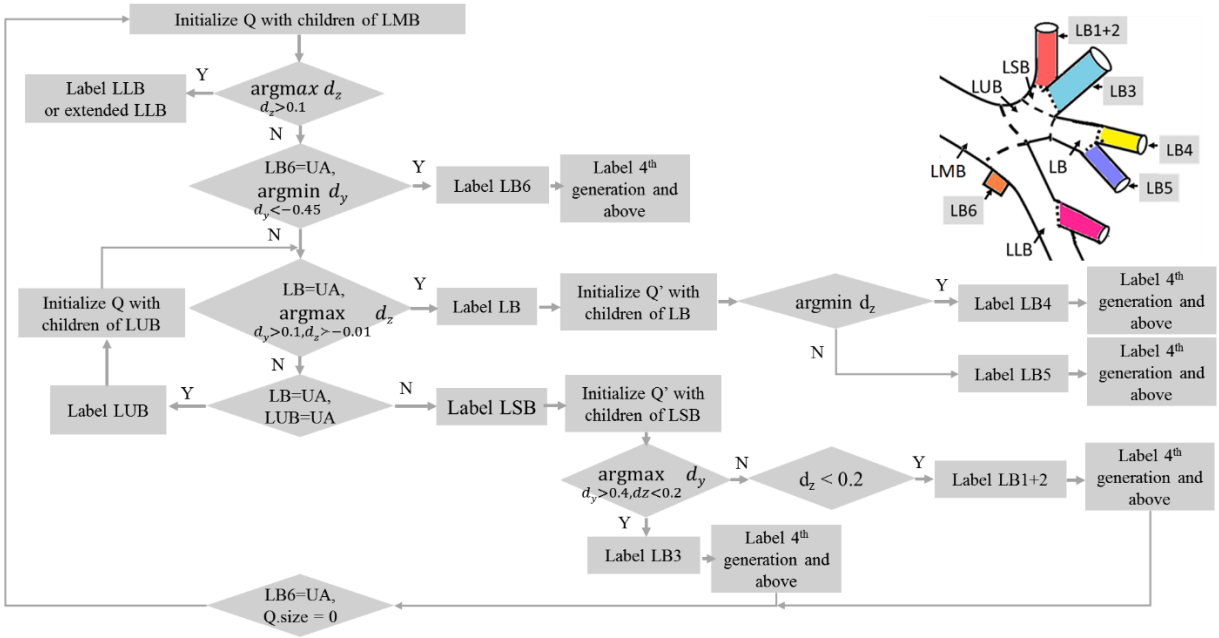

Figure 4.8. Flow chart algorithm of labeling LB1 to LB6. Unassigned label is denoted by UA.

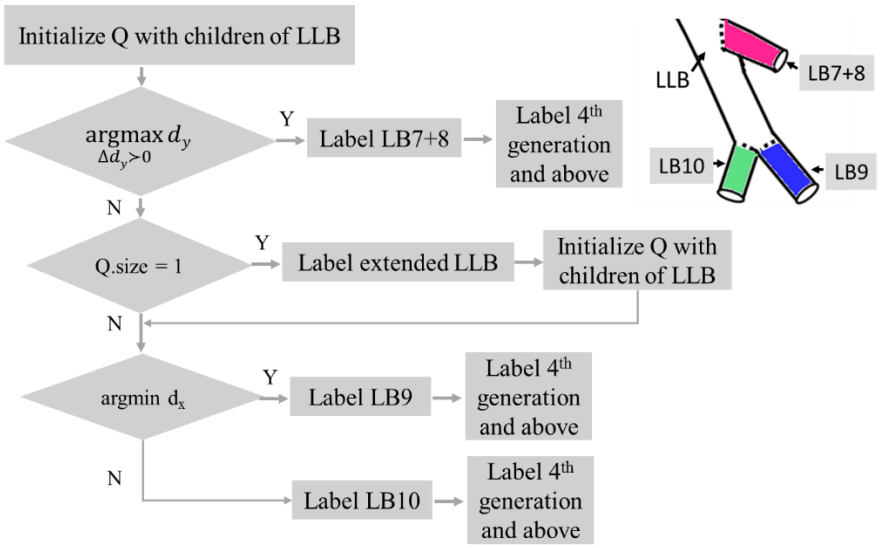

Figure 4.9. Flow chart algorithm of labeling LB7+8 to LB10.

4.2 Experiments

Two experiments were conducted to validate the hypothesis that the framework generates correct airway segmentation and anatomical labeling, and that repeatable measurements of the bronchi were achieved. First, the accuracy of the airway segmentation and labeling were evaluated by visual inspection of customized 3D color visualizations of the label outcomes for 2727 CT scans. Second, the repeatability of the bronchi dimension measurements was evaluated by comparing the measurement results from longitudinally separated scans where the assumption was made that any change in actual airway dimensions would, in general, be very small. In this case 504 CT scan pairs were compared.

4.2.1 Label accuracy experiment

The fully automated quantitative airway biomarker analysis framework was applied to a heterogeneous dataset of 2727 LDCT scans. These scans were acquired in several institutions with various CT scanner manufactures, scanner models, reconstruction kernels

and slice thickness (< =2.0 mm). To evaluate the airway segmentation and anatomical labeling performance, 3D visualizations of the airway segmentation and anatomical labeling results were generated for each scan. One coronal view and two sagittal views were employed for the visual inspections using 3D rendering for all color labeled bronchi and including the lung regions for context as shown in Figure 4.10. The trachea (zeroth generation) was colored by cyan, two main bronchi (first generation) were colored by purple and the second-generation bronchi were colored by beige, as listed in Table 1. Each of the 18 segmental level (third generation) bronchi, as listed in Table 1, was colored so that nearby segmental bronchi do not have the same color. Bronchi of fourth and above generations were colored with the same color of their third-generation ancestors.

The evaluation results were classified into two categories: Good and unacceptable. Good results should not contain any visible airway segmentation error (such as over-segmentation or under-segmentation) or wrong anatomical labeling. Otherwise, the result is considered as unacceptable. A case with good segmentation and labeling is shown in Figure 4.10; A case with unacceptable anatomical label is shown in Figure 4.11, in which LB1+2 and LB3 are mislabeled as indicated by arrows.

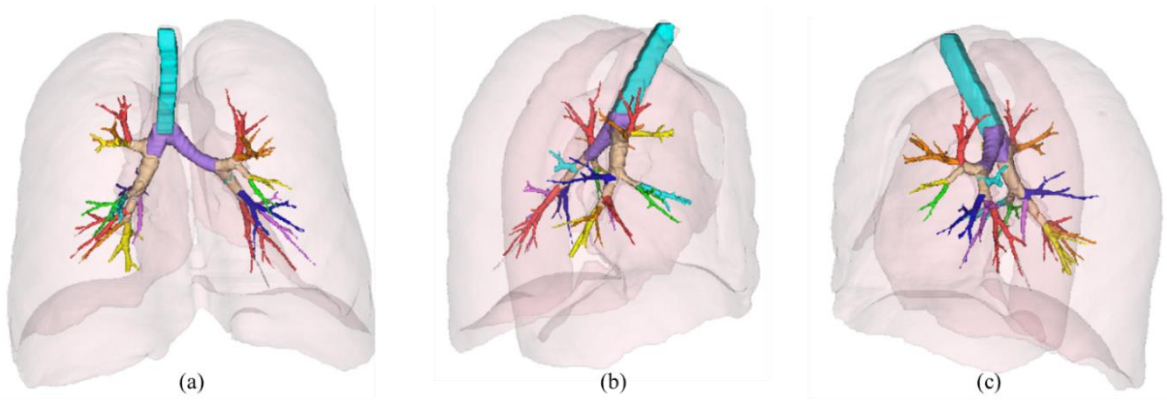

Figure 4.10. 3D visualizations of a good airway segmentation and labeling result. (a) Coronal view. (b) Sagittal view (from right to left). (c) Sagittal view (from left to right). The segmented lungs are also shown as the reference. The color legend is listed in Table 4.1.

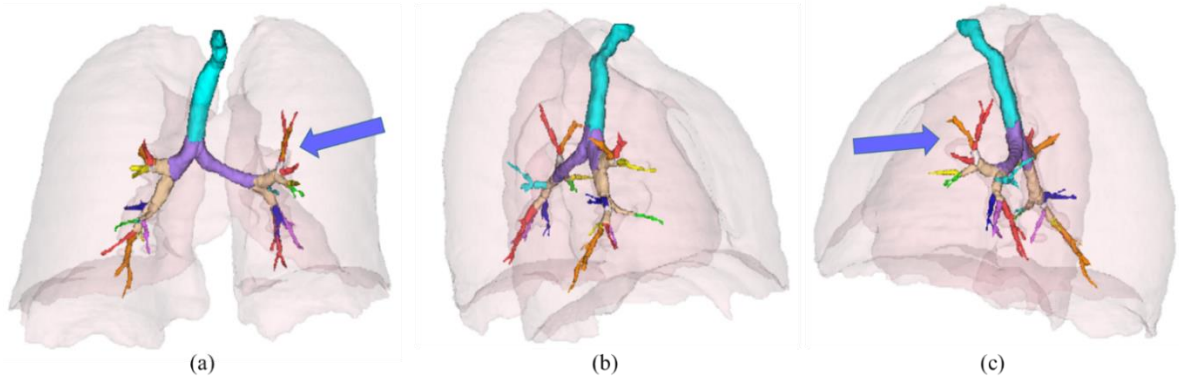

Figure 4.11. 3D visualizations of an unacceptable airway segmentation and labeling result. LB1+2 and LB3 are mis-labeled as indicated by the error. (a) Coronal view. (b) Sagittal view (from right to left). (c) Sagittal view (from left to right). The segmented lungs are also shown as the reference. The color legend is listed in Table 4.1.

4.2.1 Bronchus measurement repeatability experiment

A subset of the above dataset, consisting of 504 cases, each of which has two longitude LDCT scans, is used for the performance evaluation of the airway dimension (lumen diameter and wall thickness) measurements. The two longitude LDCT scans of the same subject were acquired at two different time-points with the assumption that there is no change occurred in airway dimensions within this interval. The performance of the airway

measurements is evaluated based on the reproducibility of the measurements, which is quantified by the percentage difference defined as following:

$$\Delta t = 100\% \times |t_{t1} - t_{t2}| / (t_{t1} + t_{t2}) \quad (4.3)$$

$$\Delta d = 100\% \times |d_{t1} - d_{t2}| / (d_{t1} + d_{t2}) \quad (4.4)$$

Where Δt and Δd are the percentage difference for the wall thickness and lumen diameter respective; t_{t1} and t_{t2} are the wall thickness measurement at time point t1 and t2 respectively; d_{t1} and d_{t2} are the lumen diameter measurement at time point t1 and t2 respectively.

A subset of above, a longitudinal dataset consisting of 86 pairs of CT scans, was used to further evaluate the impact of the inspiration level difference and time interval of the scan pairs on the reproducibility of the airway dimension (lumen diameter and wall thickness) measurements. The two longitude LDCT scans of each subject were acquired with time interval < 1 year and at similar inspiration level, which is defined by relative percentage lung volume difference < 10%.

4.3 Results

89.4% (2439 out of 2727) scans have good airway segmentation and anatomical labeling based on visual inspection. 5.0% (135 out of 2727) scans have unacceptable airway segmentation and 5.6% (153 out of 2727) scans have good airway segmentation but unacceptable anatomical airway labeling. The number of labeled bronchi of each anatomical generation (generation 0 to 9) is shown in Table 4.2. The number of each labeled segmental level bronchus (3^{rd} anatomical generation) is shown in Table 4.3.

Table 4.2. The number of labeled bronchi of each anatomical generation.

Generation	Count	Percentage
0	2470	90.5%
1	4944	90.6%
2	17971	82.4%
3	27051	55.1%
4	22636	
5	9930	
6	3305	
7	994	
8	353	
9	217	

Table 4.3. The number of each labeled segmental level bronchus (3rd anatomical generation).

Segmental level bronchus	Count	Percentage
RB1	1995	73.16%
RB2	1410	51.71%
RB3	1636	59.99%
RB4	1434	52.59%
RB5	1417	51.96%
RB6	1776	65.13%
RB7	1214	44.52%
RB8	1622	59.48%
RB9	1147	42.06%
RB10	1147	42.06%
LB1+2	1812	66.45%
LB3	1769	64.87%
LB4	1068	39.16%
LB5	1048	38.43%
LB6	2092	76.71%
LB7+8	1794	65.79%
LB9	1335	48.95%
LB10	1335	48.95%

The evaluation results of the airway dimension measurements for the longitudinal datasets are shown in Tables 4.4 – 4.7. For each case, the percentage difference of lumen diameter measurement Δd and the percentage difference of wall thickness Δt between the two

longitudinal scans are computed. The average, min, max and standard deviation of the percentage difference are reported. The reproducibility of airway dimension measurements at each anatomical generation is summarized in Table 4.4 for the longitudinal dataset of 504 cases, and in Table 4.6 for the longitudinal dataset of 86 cases. The reproducibility of airway dimension measurements at each segmental level bronchus is summarized in Table 4.5 for the longitudinal dataset of 504 cases, and in Table 4.7 for the longitudinal dataset of 86 cases.

Table 4.4. Reproducibility of airway dimension measurements at each anatomical generation for the longitudinal dataset of 504 cases. Δd_n is the percentage difference of lumen diameter measurement at the n-th anatomical generation. Δt_n is the percentage difference of wall thickness measurement at the n-th anatomical generation. The average, min, max and standard deviation (SD) of the percentage differences are shown.

Measurements	count	Average %	Min %	Max %	SD
Δd_0	504	3.18	0.00	25.78	3.40
Δt_0	504				
Δd_1	504	4.09	0.00	23.27	4.11
Δt_1	504				
Δd_2	504	8.22	0.00	33.06	6.32
Δt_2	504	9.15	0.00	44.62	7.77
Δd_3	467	9.40	0.00	56.12	7.61
Δt_3	467	10.57	0.00	57.14	9.16
Δd_4	351	10.44	0.00	78.81	9.93
Δt_4	351	9.39	0.00	43.05	7.47
Δd_5	161	13.91	0.28	89.31	13.83
Δt_5	161	9.61	0.00	54.27	8.90
Δd_6	59	19.07	0.27	78.11	18.87
Δt_6	59	9.74	0.00	31.78	7.15
Δd_7	14	18.85	0.26	59.30	18.29
Δt_7	14	12.25	2.11	30.67	8.39

Table 4.5. Reproducibility of airway dimension measurements at each segmental level bronchus for the longitudinal dataset of 504 cases. Δd_n is percentage difference of lumen diameter measurement at the segmental level bronchus n. Δt_n is percentage difference of wall thickness measurement at the segmental level bronchus n. The average, min, max and standard deviation (SD) of the percentage differences are shown.

Segmental level bronchus	count	Average%	Min %	Max %	SD
Δd_{RB1}	366	16.63	0.00	86.74	16.27
Δt_{RB1}	366	15.56	0.00	100.85	14.41
Δd_{RB2}	232	15.50	0.00	63.99	13.03
Δt_{RB2}	232	11.13	0.00	46.30	8.60
Δd_{RB3}	283	16.04	0.17	63.60	15.50
Δt_{RB3}	283	17.91	0.00	99.82	18.11
Δd_{RB4}	259	16.35	0.00	64.97	13.61
Δt_{RB4}	259	13.48	0.00	94.77	15.49
Δd_{RB5}	256	15.38	0.00	80.85	14.61
Δt_{RB5}	256	15.76	0.00	85.95	15.81
Δd_{RB6}	322	17.64	0.00	88.89	15.95
Δt_{RB6}	322	13.35	0.00	80.69	12.08
Δd_{RB7}	214	14.02	0.00	60.30	11.83
Δt_{RB7}	214	12.76	0.00	68.75	12.02
Δd_{RB8}	239	16.74	0.00	79.45	15.75
Δt_{RB8}	239	11.78	0.00	79.03	11.40
Δd_{RB9}	133	20.73	0.00	95.22	19.53
Δt_{RB9}	133	14.57	0.63	59.69	12.01
Δd_{RB10}	133	21.98	0.00	76.01	18.65
Δt_{RB10}	133	12.23	0.00	79.25	13.12
Δd_{LB1+2}	300	16.81	0.18	80.73	15.36
Δt_{LB1+2}	300	12.58	0.00	76.56	12.46
Δd_{LB3}	299	17.98	0.00	76.78	16.68
Δt_{LB3}	299	12.52	0.00	85.71	13.80
Δd_{LB4}	147	16.88	0.00	72.83	14.85

Δt_{LB4}	147	11.71	0.51	44.95	8.98
Δd_{LB5}	142	13.56	0.00	67.79	13.18
Δt_{LB5}	142	12.39	0.55	72.19	11.65
Δd_{LB6}	397	15.25	0.00	71.79	13.74
Δt_{LB6}	397	13.95	0.00	67.69	13.34
Δd_{LB7+8}	316	19.58	0.25	72.76	15.55
Δt_{LB7+8}	316	15.65	0.50	78.30	14.56
Δd_{LB9}	205	19.35	0.00	88.03	17.62
Δt_{LB9}	205	18.31	0.00	96.88	17.94
Δd_{LB10}	205	18.00	0.00	86.05	18.03
Δt_{LB10}	205	10.94	0.00	73.09	9.70

Table 4.6. Reproducibility of airway dimension measurements at each anatomical generation for the longitudinal dataset of 86 cases. Δd_n is percentage difference of lumen diameter measurement at the n-th anatomical generation. Δt_n is percentage difference of wall thickness measurement at the n-th anatomical generation. The average, min, max and standard deviation (SD) of the percentage differences are shown.

Measurements	count	Average %	Min %	Max %	SD
Δd_0	86	2.07	0.00	11.07	1.84
Δt_0	86				
Δd_1	86	2.50	0.00	20.13	2.63
Δt_1	86				
Δd_2	86	6.99	0.26	28.38	5.38
Δt_2	86	8.66	0.00	44.62	8.38
Δd_3	81	8.89	0.20	56.12	9.04
Δt_3	81	9.78	0.00	47.84	8.96
Δd_4	62	9.28	0.00	68.79	10.46
Δt_4	62	8.74	0.55	36.95	7.94
Δd_5	38	13.97	0.79	89.31	15.19
Δt_5	38	9.47	0.00	54.27	12.09
Δd_6	12	21.86	2.21	78.11	24.69
Δt_6	12	7.05	0.00	25.22	6.88

Δd_7	2	7.54	4.67	10.40	2.87
Δt_7	2	14.05	7.85	20.25	6.20

Table 4.7. Reproducibility of airway dimension measurements at each segmental level bronchus for the longitudinal dataset of 86 cases. Δd_n is percentage difference of lumen diameter measurement at the segmental level bronchus n. Δt_n is percentage difference of wall thickness measurement at the segmental level bronchus n. The average, min, max and standard deviation (SD) of the percentage differences are shown.

Segmental level bronchus	count	Average%	Min %	Max %	SD
Δd_{RB1}	65	14.79	0.00	70.89	16.81
Δt_{RB1}	65	12.14	0.00	50.98	10.71
Δd_{RB2}	43	11.97	0.27	54.02	10.61
Δt_{RB2}	43	11.85	0.61	46.30	10.55
Δd_{RB3}	54	17.84	0.27	58.20	15.80
Δt_{RB3}	54	21.28	0.49	90.55	22.47
Δd_{RB4}	50	17.56	0.26	48.65	12.79
Δt_{RB4}	50	13.17	0.00	71.22	15.93
Δd_{RB5}	50	12.53	0.00	52.35	11.60
Δt_{RB5}	50	15.74	0.00	85.95	18.03
Δd_{RB6}	63	14.60	0.19	55.11	11.22
Δt_{RB6}	63	12.21	0.00	38.95	9.67
Δd_{RB7}	47	13.72	0.00	55.79	13.27
Δt_{RB7}	47	10.63	0.00	53.33	11.42
Δd_{RB8}	45	11.86	0.11	46.48	9.95
Δt_{RB8}	45	9.79	0.00	39.13	8.61
Δd_{RB9}	37	15.88	1.09	54.31	14.39
Δt_{RB9}	37	13.00	0.63	59.69	10.98
Δd_{RB10}	37	20.22	0.56	68.65	20.93
Δt_{RB10}	37	10.32	0.00	53.41	11.68
Δd_{LB1+2}	55	16.81	0.71	64.00	16.39

Δt_{LB1+2}	55	12.31	0.00	58.14	12.91
Δd_{LB3}	55	16.85	0.13	68.03	16.83
Δt_{LB3}	55	13.97	0.58	75.35	17.60
Δd_{LB4}	24	21.51	0.29	72.83	20.59
Δt_{LB4}	24	11.52	1.99	35.22	9.54
Δd_{LB5}	21	15.07	0.68	57.72	16.20
Δt_{LB5}	21	14.79	0.69	61.89	14.61
Δd_{LB6}	74	14.84	0.00	61.00	13.36
Δt_{LB6}	74	13.15	0.00	57.56	13.51
Δd_{LB7+8}	63	15.96	0.72	54.08	14.03
Δt_{LB7+8}	63	14.07	0.50	72.27	14.10
Δd_{LB9}	48	20.03	0.28	88.03	19.31
Δt_{LB9}	48	14.26	0.66	47.89	12.00
Δd_{LB10}	48	16.91	0.19	56.67	16.27
Δt_{LB10}	48	8.74	0.00	38.40	7.94

4.4 Discussion

The number of labeled bronchi is summarized in Tables 4.2 and 4.3. The percentage of the labeled bronchi decreases as the anatomical generation increases, which can be explained by the fact that the more peripheral airway bronchus, often with smaller lumen diameter, is more difficult to segment or measure accurately given the limited scan resolution. There is a large variation in the percentage of the labeled bronchi at the segmental level (3^{rd} anatomical generation), as shown in Table 4.3. More than 80% of RB1 and LB6 can be obtained; while only approximately 40% of RB9, RB10, LB4 and LB5 can be obtained. This can also be explained by the fact that bronchi with smaller lumen diameter are more challenging to segment or measure accurately as different segmental level bronchi are located at different lung regions and have different lumen diameters. For example, RB9 and RB10 are considered

to be more challenging to obtain, as they usually have smaller lumen diameter compared to RB1.

The reproducibility of airway dimension measurements for the longitudinal datasets is summarized in Tables 4.4-4.7. A percentage difference of less than 10% was obtained for most segmental level (3rd generation) bronchi as indicated in Tables 4.4 and 4.6. Whether this reproducibility of measurement is sufficient to be clinically useful? The answer depends on the impact of diseases of interest on the airway dimensions, which is still a subject of ongoing research.

The measurement reproducibility (as measured by the percentage difference in longitudinal scans) decreases as the anatomical generation increases, as shown in Tables 4.4 and 4.6, which can be explained by the fact that more peripheral bronchi (with smaller lumen diameter) are more difficult to measure accurately given the limited scan resolution. In addition, the more peripheral airway dimensions are more likely to be affected by the airway diseases [101, 114, 103] and by the inspiration levels [118], leading to a larger percentage difference in longitudinal scans. The measurement reproducibility varies across the segmental level bronchi and the wall thickness in general has better reproducibility compared to lumen diameter, as shown in Tables 4.5 and 4.7. For the dataset of 504 cases as shown in Table 4.5, the most robust wall thickness measurement is obtained at RB2 with a percentage difference of 11.13% and the most robust lumen diameter measurement is obtained at LB5 with percentage difference of 13.56%. For the dataset of 86 cases as shown in Table 4.7, the most robust wall thickness measurement is obtained at RB8 with percentage difference of 9.79% and the most robust lumen diameter measurement is obtained at RB8 with percentage difference of 11.86%. The differences in the measurement robustness across the segmental

level bronchi suggest that certain anatomical bronchi may be more suitable to be used in the study of the airway dimensions.

The measurement reproducibility of the 86-case dataset with more strictly controlled inspiration level difference and scan time interval, is better than that of the 504-case dataset, as shown in Tables 4.4-4.7. This can be explained by the fact that in the reproducibility study using longitudinal datasets, the validity of the assumption that there is no change occurred in airway dimensions within the time interval between the two longitudinal scans is influenced by two main factors: the length of the time interval and the inspiration level difference. If the time interval between the two longitudinal scans is long, airway diseases are more likely to develop, which may lead to variations in the airway dimensions. The inspiration level also has a huge impact on the airway dimensions as pointed out by Petersen et al. in [118]: Airway lumen diameter increases and wall thickness decreases with inspiration. The effect of inspiration is greater in higher-generation airways. The experiment results of the measurement reproducibility in the two longitudinal datasets suggest that for the future work, the inspiration level difference and scan time interval of the longitudinal scan pairs must be carefully controlled to achieve precise airway measurements and meaningful longitudinal comparison.

4.5 Conclusion

A fully automated anatomy directed framework is presented for the analysis of reproducible quantitative airway biomarkers from LDCT scans acquired from the annual lung cancer screening. The airway is first segmented with each bronchus labeled with corresponding anatomical name. The lumen diameter and wall thickness of each bronchus are then measured at each bronchus, serving as reproducible biomarkers that provide valuable information aiding the diagnosis and treatment of COPD. The framework was evaluated with

a dataset of 2727 LDCT scans and good airway segmentation and labeling were obtained in 89.4% of scans based on visual inspection. The airway biomarker measurement is evaluated with a longitudinal dataset of 504 cases and demonstrates good reproducibility in both lumen diameter measurement and wall thickness measurement.

CHAPTER 5

PULMONARY NODULE CLASSIFICATION USING 3D CONVOLUTIONAL NEURAL NETWORK

Annual lung cancer screening with low-dose chest CT has recently been approved in the United States for the early detection and treatment of lung cancer for people at high risk, with approximately 8.7 million Americans eligible for the screening [121]. The costly follow-up procedures provide motivation for the development of systems for establishing the malignancy status of pulmonary nodules from low-dose chest CT images. The purpose of this study is to determine the benefits of applying a novel machine learning approach, 3D convolution neural network (CNN), to the task of pulmonary nodule classification from low-dose chest CT scans obtained from lung cancer screening, through the performance comparison with traditional machine learning approaches. In addition, the classifier ensembles of the different combinations of the 3D CNN and traditional machine learning classifiers based on handcrafted 3D image features are also explored to study the key to the success of the ensembles.

A typical automated system for lung cancer diagnosis generally consists of two stages: pulmonary nodule detection and pulmonary nodule malignancy classification. This study focuses on the latter stage, namely the discrimination between benign pulmonary nodules and malignant pulmonary nodules given the nodule location and size from the volumetric low-dose chest CT scans acquired during the lung cancer screening.

The conventional automated approaches to the discrimination between benign pulmonary nodules and lung cancer generally consist of four major stages [122]: 1) Nodule segmentation; 2) Image feature extraction from the segmented nodules; 3) Feature selection based on the discriminative power of the features; and 4) Machine learning classifier training given the selected features. A wide range of image features, such as gray-level distribution, size, morphology as well as texture description, and various types of machine learning models, including linear discriminant analysis [123, 124], support vector machines (SVM) [125, 126], massive training artificial neural network [127], random forest [122], and distance weighted nearest neighbor [126, 128], have been explored in the literature [122, 129, 124, 125, 126, 128] to address the problem of pulmonary nodule classification. The fast volume growth rate of a nodule [23] serves as a reliable indicator for malignancy, however it usually requires more accurate image segmentation for the nodule volume measurement and at least two CT scans, which prolongs diagnosis and exposes the patient to possibly unnecessary radiation exposure [125, 126, 128].

The astounding revival of convolutional neural networks (CNNs) [130, 131] since 2012, owing to the availability of large-scale annotated image datasets [132] and affordable parallel computing resources [133], has led to remarkable advances [131, 134, 135, 136] in several computer vision applications of natural images and the birth of deep learning, a new area of machine learning research. The application of deep learning techniques to various automated medical imaging analysis problems has also been explored in a large number of published work [137], which can be summarized in the following three primary categories [138]. First, off-the-shelf deep features can be extracted directly from pre-trained deep learning networks and then fed into traditional machine learning models, such as SVM and

random forest, to address the detection or classification problems [139, 140, 141, 142]. Although the performance of using only off-the-shelf deep features is normally inferior to the traditional state-of-the-art features acquired by careful feature engineering [142, 139, 140, 141], the ensemble of both can result in substantial improvement [139, 140, 141]. Second, fine-tuning deep learning models pre-trained on irrelevant and typically non-medical images, has been demonstrated to outperform the state-of-the-art traditional approaches [143, 138]. Third, effective deep learning models can also be trained from scratch. It can be used either from-to-end [133, 138, 144, 145, 146, 147, 148, 149, 150, 151] [152] or as a feature extractor only [153, 145, 154, 155], which requires succeeding traditional approaches such as the conventional classifier for classification applications or the deformable shape model for segmentation applications.

Pulmonary nodule detection and classification from CT scans is a 3D problem, whereas most of the published work on deep learning still adopts a 2D approach [133, 156], since the CNNs were originally proposed for 2D natural images with RGB color channels. In order to utilize the established network architectures and pre-trained network weights, the most common solution is to map each 3D volumetric CT scan into a 3-channel 2D image by assigning 3 orthogonal planes, which can be axial, coronal and sagittal slices [139, 142, 143] or even planes with random orientations [133, 145, 144], to 3 different channels. It effectively reduces the network complexity, in terms of the number of trainable weights and the required memory for the computation as well as data storage, and the amount of the training data needed to avoid overfitting; however, the concern of the loss of 3D information still exists [156]. A number of recent publications have started to employ 3D CNNs in various types of medical imaging applications, including pulmonary nodule [147] and cerebral microbleeds

[148] detection, prostate finding [149] and breast mass [150] classification, and different types of anatomy segmentation [151, 152, 157, 154, 155]. 3D CNN has been shown to achieve significant performance enhancement attributed to the consideration of contextual information along the 3rd spatial dimension, compared to the corresponding 2D CNN, by Cicek et al. [151] for the segmentation of xenopus kidney, by Dou et al. [148] for the detection of cerebral microbleeds, and by Li et al. [150] for the classification of breast masses. For the application of pulmonary nodule classification, no published work has been found on the employment of a 3D CNN, which may potentially provide benefits due to the consideration of the full 3D data [125]. In the CNN proposed by Shen et al [146, 145], although 3D image patches around the nodules are directly fed into the input layer, the network is still not completely 3D, because no convolution operation or pooling operation is performed along the 3rd spatial dimension, which is treated as the channel dimension.

The matching of size distribution for malignant and benign nodules in the validation set is necessary for a meaningful assessment of automated systems for pulmonary nodule classification as first noted in 2007 [128] and also in [122, 126]. Datasets with benign nodules dominating the small size range and malignant nodules dominating the large size range are very common in the published studies [158, 123, 124, 127, 159], since the nodule malignancy is highly correlated to nodule size. However, algorithm performance evaluated on such a dataset can be misleading and overly optimistic [122, 126, 128], because a simple size classifier that is based on nodule size thresholding only, may achieve promising performance, due to correctly classifying very large and very small nodules. However, such a classifier would not be effective to classify the malignancy status of nodules of intermediate sizes which are the most frequent in lung cancer screening and of most interest to clinical practice.

A balanced class distribution (i.e., approximately equal number of benign and malignant nodules in our case) is another favorable property of the validation set for a classification problem evaluated using receiver operating characteristic (ROC) curves. A large skew in class distribution in the validation set can lead to overly optimistic view of an algorithm's performance based on ROC curves [160]. Therefore, it can be unfair to directly compare the ROC curves of the algorithms evaluated on the datasets with different amount of skewness.

The reported performance of automated nodule classification systems spans a very large range [122, 123, 124, 125, 23, 127, 128, 143, 153, 145] [146], with area under the ROC curve (AUC) ranging from 0.50 [122] to 0.93 [146]. However, the performance between studies is generally not comparable due to two primary reasons [122]. First, different datasets were employed for the evaluation of each study; therefore, direct comparison is meaningless [161]. And especially as discussed above, a biased validation set may lead to unfair assessment of its performance [138]. Moreover, studies [123, 124, 126, 127, 128] that focus on pulmonary nodule classification from low-dose CT images acquired during lung cancer screening are considered more challenging [126, 128] compared to other studies [122, 125, 143, 153, 145, 146] that include standard-dose CT images acquired during clinical practice, due to the small size of the present nodules and the high level of image noise [23, 127]. The current scan protocol in lung cancer screening (fixed CT scan resolution of 512 pixels across lungs) limits the number of pixels available for analysis especially for nodules of small size, which are the most clinically relevant for early detection of lung cancer. Second, different evaluation schemes were used. For instance, the malignancy status is confirmed by biopsy outcome in some studies [122, 123, 124, 126, 127, 128, 153], whereas in other studies [125,

143, 145, 146], the malignancy status is established purely based on malignancy ratings of radiologists after reviewing the CT scans, where inter-observer differences can be significant [122, 143]. In addition, the cross-validation strategy, such as leave-one-out compared to 5-fold-cross-validation, may have a significant impact on the resulting performance [138].

In this study, we have applied a 3D CNN trained from scratch to the classification of pulmonary nodule malignancy using a class-balanced and size-matched low-dose chest CT dataset, where the malignancy status is pathologically confirmed. Since the exact same training and validation dataset as well as the evaluation scheme were employed in a previous study based on handcrafted features and traditional machine learning models by Reeves et al [126], a direct performance comparison between the 3D CNN and conventional approaches to the pulmonary nodule classification is possible. The ensemble models of the different combinations of the 3D CNN and traditional machine learning models were also explored. Moreover, the performance comparison between 3D CNN and 2D CNN architectures were investigated. Our hypothesis is that the CNN can learn the 3D image features automatically and achieve at least the same classification performance compared to the conventional approaches that are based on handcrafted image features and traditional machine learning classifiers. In addition, since the features learned by CNN should be complementary to the handcrafted features, the ensembles should achieve further performance improvement. Finally, performance enhancement can be obtained by using 3D CNN compared to using 2D CNN due to the consideration of the of contextual information along the 3rd spatial dimension.

5.1 3D Convolutional neural network architectures

A CNN [130] is a specialized type of feedforward neural network (or multilayer perceptrons), which incorporates convolution operations in at least one of its computational layers and is typically applied to input data with grid-like topology, such as image data [162]. A feedforward neural network is made up by a number of concatenated computational layers, where the computational outcome, namely feature map, of each layer is simply a mathematical mapping of the output of the previous layer. The composition of all computational layers contained in the network together defines a mapping Y = f (X; θ) from the input tensor X (3D image matrix in our study here) to the output tensor Y (1D class vector in our study here), where θ is a set of mapping parameters (or weights) to be learned during the training process [162].

A Conv layer maps an input tensor X to an output tensor Y by the convolution of kernel tensor K and the input X across each spatial axis. Let $Y_{c,\,x,\,y,\,z}$ denote value of the c-th channel of the output 4D tensor Y (namely 4D feature map) at spatial location (x, y, z), then the computation defined by a 3D Conv layer is given in equation (5.1):

$$Y_{c,x,y,z} = Conv(X,K) + B_c = \sum_{c'} \sum_{i,j,k} X_{c',\ xs+i,\ ys+j,\ xs+k}\ K_{c,c',i,j,k} + B_c \qquad (5.1)$$

Where s is the stride for convolution; $X_{c',x,y,z}$ is the value of c'-th channel of the input 4D tensor X (usually feature map output by previous layer if it is not the input data layer) at spatial location (x, y, z); B_c is the bias for the c-th output channel; and $K_{c,c',i,j,k}$ is the (i, j, k)-th element of the kernel, corresponding to the connection strength between a unit in the c'-th input channel and a unit in the c-th output channel with spatial offset of (i, j, k) between the

output unit and the input unit [162]. The summation iterates over all input channels and the 3D spatial dimensions of the kernel.

A FC layer defines a mapping that is simply an inner product of the input vector X and weight matrix W with the addition of the bias vector B. Thus, the output vector Y can be computed by equation (5.2):

$$Y = W X + B \tag{5.2}$$

The ReLU layer is employed to incorporate non-linearity into the network and thus to increase the capacity of the overall model. It is an element-wise operation defined for each input unit x as shown in equation (5.3):

$$y = \max(0, x) \tag{5.3}$$

The Softmax layer and cross entropy loss layer are generally located at the end of the network and together serve as the output layer. The Softmax layer is used for converting the input N-dimensional (i.e., N classes) prediction score vector Y into an N-dimensional prediction probability vector ranging from 0 to 1. If we let P_c denote the output probability for class c, and Y_c denote the input prediction score for class c, then the Softmax function is defined in equation (5.4):

$$P_c = \text{Softmax}(Y)_c = \frac{e^{Y_c}}{\sum_{d=1}^{N} e^{Y_d}} \tag{5.4}$$

The cross entropy loss layer takes the ground truth class label c' and prediction probability for class c' to compute the cross entropy loss L as shown in equation (5.5):

$$L = - \log P_{c'} \tag{5.5}$$

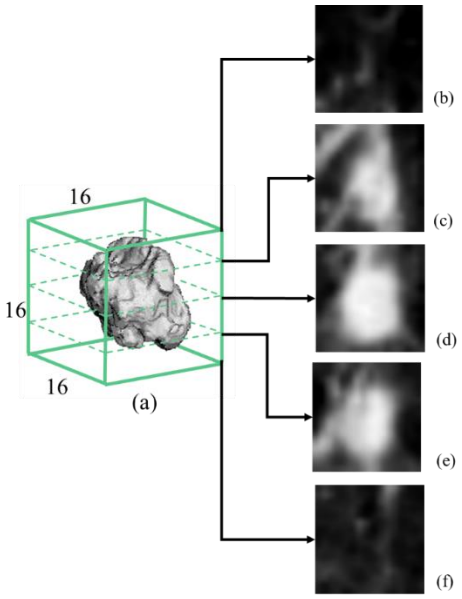

Figure 5.1 Cropped 3D CT volume used as the input to the CNN. (a). 16x16x16 re-sampled isotropic CT volume centered at a nodule. (b-f) 5 axial slices at the corresponding axial level.

5.1.1 Nodule image pre-processing

Each nodule CT volume is cropped into a real space cube around the nodule center with a margin of 20% of the nodule radius as shown in Figure 5.1, to include approximately the same volume for background context as the nodule itself. The nodule location (center) and size (radius) are determined from automated nodule segmentation [126, 23]. The cropped 3D image region is then resampled using tricubic interpolation to an isotropic fixed size 3D image.

For these CT images, the x and y dimensions have the same resolution and the z dimension (slice spacing) usually has a lower resolution. As detailed in section 5.2.1, the pixels in the data set, in general, varied in a size range from 0.5 to 0.85 mm in the x and y dimensions and from 1.0 to 2.5 mm in the z dimension. Two different resampled image sizes were independently explored for the CNN network: 16x16x 16 and 32x32x32. For the

32x32x32 image size, oversampling occurred in all three dimensions since none of the cropped image regions had any dimension with more than 32 pixels in the original CT scan. For the 16x16x16 image size, 27.6% of the cropped image regions (corresponding to the largest nodules) had an xy dimension greater than 16 pixels; no cases had more than 16 pixels in the z dimension (the largest cropped region xy dimension was 31 pixels (median = 13) and the largest z dimension was 15 pixels). Therefore, for these 27.6% cases there was some amount of undersampling (possible information loss), although oversampling always occurred in the z dimension.

The image pixel values are first converted to the Hounsfield unit (HU) scale, then clipped between [-800, 200] HU considering the common image intensity distribution of pulmonary nodules, and scaled by 1/200. The resulting image intensity distribution is in the range [-4, 1] with most of the nodule pixels approximately zero-centered in the range [-1, 1].

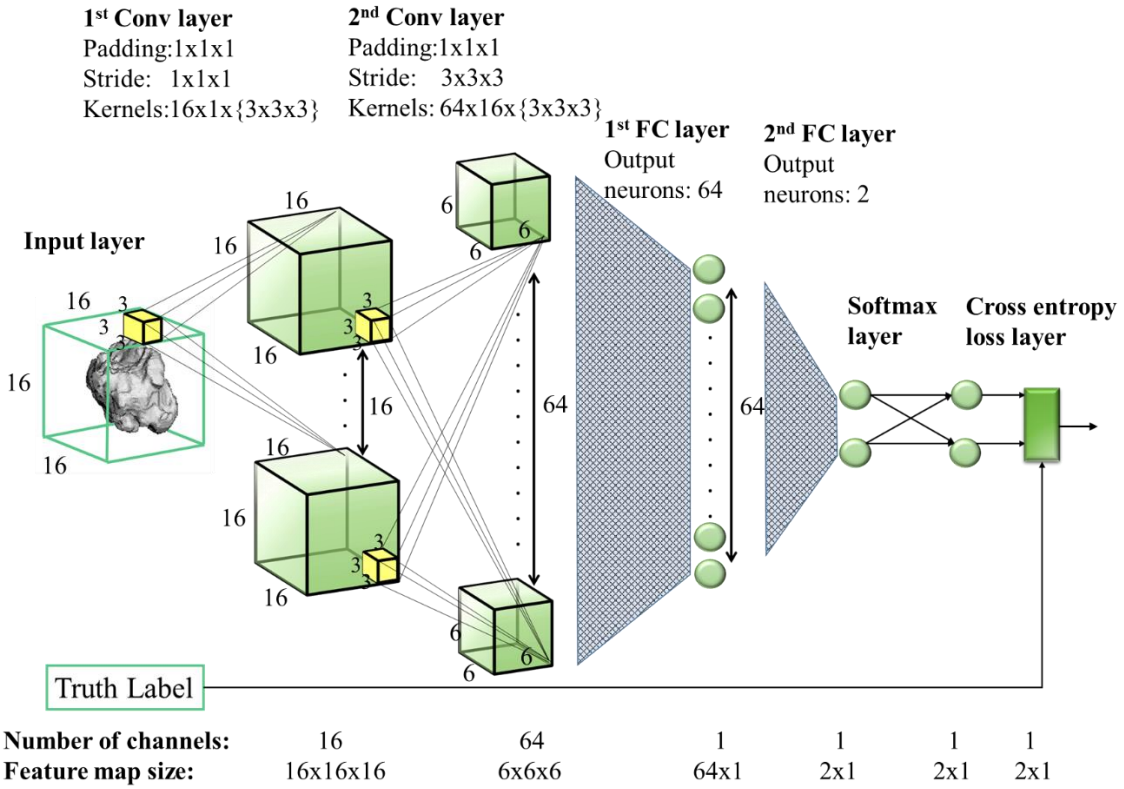

Figure 5.2. The presented CNN1 architecture. The spatial dimension and the number of channels of the feature map in each hidden layer are denoted on the bottom. The dimension and the number of kernels as well as the size of padding and stride used in each Conv layer and the number of output neurons in each FC layer are denoted on the top.

5.1.2 3D CNN architectures

Three 3D CNN architectures, CNN1, CNN2 and CNN3, are considered in this study. CNN1 takes an input image of size 16x16x16, and consists of two 3D convolutional (Conv) layers followed by two fully connected (FC) layers with one Rectified Linear Units (ReLU) layer inserted between each pair of adjacent hidden layers, as illustrated in Figure 5.2. The spatial dimension and the number of channels of the feature map in each hidden layer are denoted on the bottom of the figure. The dimension and the number of kernels as well as the size of padding and stride used in each Conv layer and the number of output neurons in each FC layer are denoted on the top of the figure. CNN2 and CNN3 employs deeper (one

additional convolutional layer) and wider network compared to CNN1; in addition, CNN2 takes larger input images of size 32x32x32 as detailed in Table 5.1.

Table 5.1. Description of CNN architectures and overall performance. The kernel size, padding size, and stride size are the same for all spatial dimensions, thus only one number is specified in the table, e.g., kernel size of 3 indicates 3x3x3 for 3D network. A Conv layer with n kernels is denoted as Conv-n, with default kernel size of 3, padding of 1, and stride of 1. A FC layer with n output neurons is denoted as FC-n. A ReLU non-linearity layer is inserted after each Conv layer and the 1st FC. All parameter values other than the default are specified explicitly below. The same output layer Softmax + Cross entropy are used for all models and not shown below.

CNN models	Input	Architecture	Overall AUC ± SD
CNN1 (3D)	Size:16x16x16 Channels: 1	Conv-16 + Conv-64 (stride of 3) + FC-64 + FC-2	**0.732** ± 0.052
CNN2 (3D)	Size:32x32x32 Channels: 1	Conv-32 + Conv-64 (kernel of 4, stride of 2) + Conv-64 (stride of 3) + FC-64 + FC-2	0.698 ± 0.047
CNN3 (3D)	Size:16x16x16 Channels: 1	Conv-32 + Conv-64 + Conv-64 (stride of 3) + FC-64 + FC-2	0.688 ± 0.048
CNN4 (2D)	Size:32x32 Channels: 32	{Conv-32 + Max-3}x3 + FC-32 + FC-2	0.712 ± 0.039
CNN5 (2D)	Size:32x32 Channels: 3	{Conv-16}x2 + Max-2 + {Conv-32}x2 + Max-2 + {Conv-64}x2 + Max-2 + FC-32 + FC-2	0.706 ± 0.041
CNN6 (2D)	Size:16x16 Channels: 16	{Conv-16 + Max-3}x3 + FC-32 + FC-2	0.706 ± 0.035
CNN7 (2D)	Size:16x16 Channels: 3	{Conv-16}x2 + Max-2 + {Conv-32}x2 + Max-2 + {Conv-64}x2 + Max-2 + FC-32 + FC-2	0.712 ± 0.056
CNN8 (2D)	Size:16x16 Channels: 16	Conv-16 + Conv-64 (stride of 3) + FC-64 + FC-2	0.679 ± 0.057
CNN9 (2D)	Size:16x16 Channels: 3	Conv-16 + Conv-64 (stride of 3) + FC-64 + FC-2	0.688 ± 0.058

The presented CNN architectures, including the input volume size, types of layers, network depth, kernel size, the number of kernels and neurons, etc., was based on the architectures proposed in several recent studies [133, 138, 146, 147, 148, 154], taking into consideration of the size of the input image volume and the training set to avoid over-fitting. It is infeasible to employ many layers of feature abstractions as discussed by Dou et. al. in [148], since our task is a binary classification problem with relatively small size of input 16x16x16 or 32x32x32. Moreover, less complicated CNN architectures are better suited due to the small scale of training dataset. In fact, in several recent studies [133, 146, 147, 148, 154] using CNN trained from scratch for discrimination task in medical imaging processing, no more than 3 convolutional layers are used and the maximal number of convolutional kernels is 64.

5.2 Ensembles of CNN and traditional models

Classifier ensembles have been shown to consistently outperform a single best classifier [139, 163, 164], assuming sufficient diversity among the included classifiers. The presented CNN model can be considered complementary to the conventional machine learning models [139, 143] because of two main reasons. First, traditional models are built upon handcrafted features that were designed empirically with respect to gray-level distribution, size, morphology as well as texture pattern; whereas the features employed by the CNN are learned by the network automatically. Second, the design of the two types of classifiers are also different, namely they target optimizing different types of loss functions with CNN potentially providing significantly increased model capacity. Therefore, ensembles of CNN and traditional models have the potential to give rise to remarkable performance enhancement.

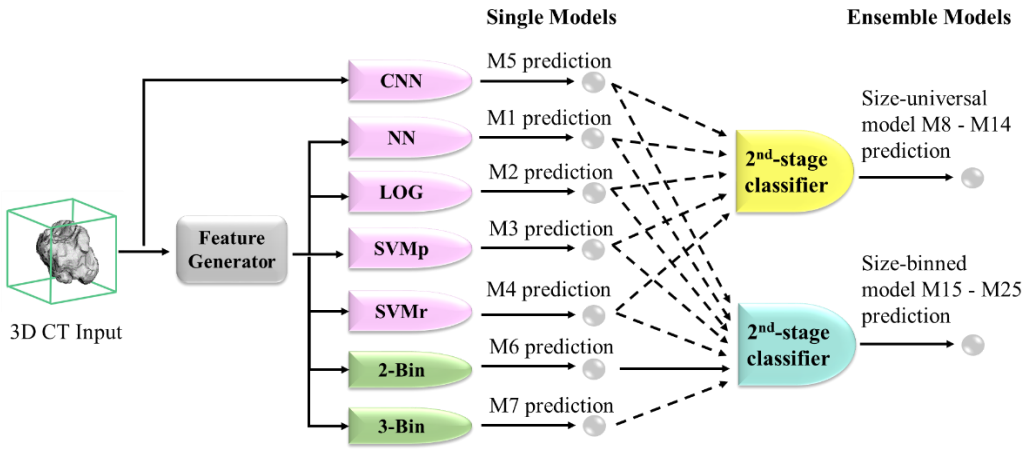

Figure 5.3. The construction of the ensemble models. Models of 4 different categories are differentiated by colors: size-universal single models (), size-binned single models (), size-universal ensemble models (), and size-binned ensemble () models. For the inputs to the 2nd-stage classifiers, the dashed line indicates the respective input may or may not be used.

Two types of traditional nodule classification models, size-universal model and size-binned model, presented by Reeves et al in [126] are used in combination with CNN model to construct ensemble models in this paper. The size-universal model consists of one classifier that is trained in the class-balanced and size-matched dataset and is applicable to classify nodules of any size. The size-binned model consists of several classifiers, each of which is trained with and applicable to nodules of a specific size range. Both 3-bin model (including B6 for diameter of (5, 7) mm, B8 for diameter of (7, 9) mm and B12 for diameter of (9,14) mm) and 2-bin model (including B6 for diameter in (5, 7) mm and B8+12 for diameter in (7, 14) mm) were presented in [126].

The same set of handcrafted image features is used in the aforementioned two types of traditional nodule classification models. The feature set consists of 46 3D image descriptors in terms of morphology, density, curvature and margin gradient. The details on the definition and generation of the image features are described by Reeves et al in [126]. Four traditional

classifiers, including distance-weighted nearest neighbor (NN) [165], logistic regression (LOG) [166], support vector machine with polynomial function kernel (SVMp) and support vector machine with radial basis function kernel (SVMr) [167], were explored for each of the two models.

The CNN model and the traditional models are combined into ensemble models by a second-stage classifier [139, 142] as illustrated in Figure 5.3. Before the combination, the classification scores predicted by each single model are first standardized to zero-mean and unit-variance. A second-stage classifier then takes the standardized scores as input features and generates classification scores to serve as the final prediction of the ensemble model.

5.3 Experiments

For the verification of the proposed hypothesis, three primary experiments were conducted. First, the three 3D CNN models and six 2D CNN models were trained and evaluated using 5-fold cross validation. Second, the 3D CNN1 model was then compared to the 4 size-universal traditional models presented by Reeves et al [126]. Since exactly the same training-validation-testing partition and evaluation scheme were employed, the effectiveness and strength of the 3D CNN model can be demonstrated. Third, the classifier ensembles constructed with different combinations of single classifiers were compared to explore the key to performance enhancement in classifier ensembles.

5.3.1 Dataset description

The dataset was constructed by combining CT scansfrom two large lung cancer screening studies, the National Lung Cancer Screening Trial (NLST) [8] and Early Lung Cancer Action Program (ELCAP) [7]. Only one instance of a nodule was used per subject.

The status of malignant nodules was confirmed by either biopsy or histology of resected tissue, while the status of benign nodules was established based on a negative outcome of the biopsy or histology of resected tissue or by 2 years of no clinical change determined by a board-certified radiologist.

The dataset is class-balanced with equal size distribution for benign and malignant nodules. The same number of benign and malignant nodules, namely 163 benign and malignant nodules, were included. The size distribution for the benign nodules is the same as that for the malignant nodules: 44.79% nodules with a diameter between 5.0 and 7.0mm, 28.22% nodules with a diameter between 7.0 and 9.0mm and 26.99% nodules with a diameter between 9.0 and 14.0mm. Only solid nodules and solid component of part-solid nodules are considered as in the study by Reeves et. al [126]. A summary of distribution of the nodule sizes and classes is given in Table 5.2.

The CT scans were obtained using a wide range of scanners, including Siemens, GE Medical Systems, Philips and Toshiba scanners, and image resolutions, where 95.4% CT scans have in-plane resolution in the range of [0.5, 1.0] mm and 98.2% CT scans have vertical resolution in the range of [1.0, 2.5] mm. More details about the process of dataset construction are described in [126].

Table 5.2. The distribution of nodule sizes and classes.

	Number of nodules	**Min size (diameter in mm)**	**Max size (diameter in mm)**	**Average size (diameter in mm)**	**Median size (diameter in mm)**
Malignant	163	5.01	14.00	8.05	7.21
Benign	163	5.02	13.91	8.01	7.27
All	326	5.01	14.00	8.03	7.22

5.3.2 Training and Testing

The dataset is randomly divided into 5 approximately equal sized folds with balanced size and class distribution. The 5-fold partition is the same with that used by Reeves et. al in [126] to ensure fair comparison. During the 5-fold cross validation, each fold is iteratively tested while the other 4 folds are further split to be used for the training set (85%) and validation set (15%). The performance of the 5 testing folds is averaged and considered as the overall performance of the testing model.

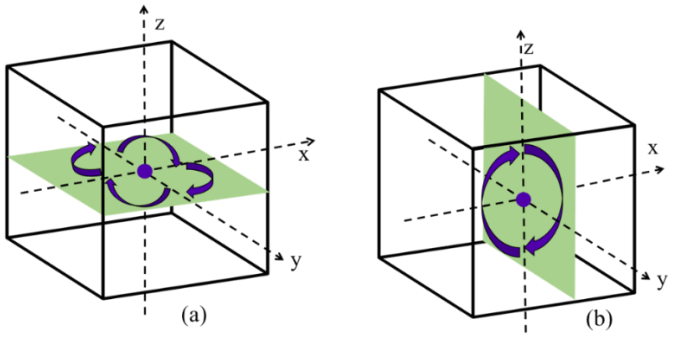

Figure 5.4. Data augmentation by rotation. (a) 4 orientations on the axial (x-y) plane. (b) 2 orientations along the vertical (z) direction.

Data augmentation is employed during the training of CNN models to reduce overfitting. Each nodule volume is rotated into 8 different orientations, including 4 rotations by 90° about the z axis and 2 rotations by 180° about the x axis (in this case, it can also be viewed as rotation about the y axis) as indicated in Figure 5.4, which results in an augmented dataset of 326 x 8 = 2608 nodule volumes. Many additional augmentations through other angle rotations or mirroring are possible. Rotations of 90° about the x or y axes were avoided due to the difference in resolution between the x, y and z dimensions; x and y resolutions of all scans are < 1.0 mm while z resolution of all scans are $\geqslant$ 1.0 mm and 60% scans are $\geqslant$ 2.0 mm. The augmented data is not used during the testing.

The weights in Conv layers are initialized using random Gaussian distributions with standard deviation of 0.01, and the weights in FC layers are initialized according to the Xavier algorithm as suggested by Glorot et al in [168]. In each round of the 5-fold cross validation, the network is trained up to 6000 epochs with the mini-batch size of 16 nodule image volumes. Early termination is adopted to reduce overfitting based on the performance tested in the validation set.

Stochastic gradient descent with moment of 0.9 is used for the training. To avoid overfitting, L2 regularization and dropout (only for the 1st FC layer) with dropout ratio of 0.5 are incorporated. The initial learning rate η_0 and weight decay C parameter are hyper-parameters tuned by random search in the range $\eta_0 \in$ [1e-5, 1e-2] and C $\in$ [1e-4, 1e-2] based on the performance tested in the validation set. The learning rate is decreased according to the strategy defined in equation (5.6):

$$\eta(n) = \frac{\eta_0}{(1+ 1e{-}4\ n)^{0.75}} \tag{5.6}$$

where $\eta(n)$ is the learning rate at training iteration of n.

The CNNs were implemented and evaluated using Caffe framework [169] on 5 Intel(R) CPUs 2.6GHZ with CentOS Linux OS and 2 NVIDIA Tesla K40c GPUs.

5.3.3 Ensembles

As summarized in Table 5.3 and Figure 5.3, 18 ensemble models are constructed using different combinations of 7 single models (M1 – M7). In order to demonstrate the benefits resulting from the incorporation of 3D CNN model into the ensembles of other traditional models, the ensemble models are constructed in pairs, one with 3D CNN and the other without 3D CNN, as in M8 vs. M9, M10 vs. M11, M12 vs. M13, M3 vs. M14, M15 vs. M16,

M17 vs. M18, M19 vs. M20, M21 vs. M22, M23 vs M24, M6 vs. M25. Different combinations of the single models are explored with gradual exclusion of models with inferior performance, such as LOG, SVMr and NN, to illustrate the effects of the number and quality of single models on the overall ensemble performance.

Table 5.3. The summary of 25 models M1 – M25. The classifiers included in each model are marked by +. Each model belongs to one of 4 different categories indicated in the leftmost column: size-universal single model (M1 – M5), size-binned single model (M6 – M7), size-universal ensemble model (M8 – M14), and size-binned ensemble model (M15 – M25). The results for each model are shown in 6 rightmost columns. The overall AUC and AUC for each size bin are averaged over 5 folds with corresponding standard deviation (σ) reported below. The right 2 columns are the p-values for the difference of ROC compared to the M3 (SVMp) and M6 (2-bin) respectively. AUC shown in bold indicates the largest in each category. P-values show in bold indicate it is below the significant level (0.05), thus statistically significant.

		Size-universal					**Size-binned**		**Results**					
		NN	LOG	SVMp	SVMr	CNN	2-bin	3-bin	AUC for B6 ± σ	AUC for B8 ± σ	AUC for B12 ± σ	Overall AUC ± σ	p-value compared to M3 (SVMp)	p-value compared to M6 (2-bin)
Size-universal Single	M1	+							0.643 ± 0.067	0.756 ± 0.032	0.769 ± 0.091	0.700 ± 0.046	0.862	0.160
	M2		+						0.458 ± 0.109	0.773 ± 0.072	0.720 ± 0.125	0.624 ± 0.090	0.289	**0.014**
	M3			+					0.620 ± 0.056	0.790 ± 0.100	0.777 ± 0.111	0.708 ± 0.056		0.494
	M4				+				0.608 ± 0.047	0.774 ± 0.098	**0.787** ± 0.089	0.699 ± 0.050	0.639	0.213
	M5					+			**0.734** ± 0.051	**0.804** ± 0.089	0.682 ± 0.119	**0.732** ± 0.052	0.256	0.444
Size-binned Single	M6						+		**0.687** ± 0.070	0.772 ± 0.103	**0.791** ± 0.064	**0.742** ± 0.049	0.494	
	M7							+	**0.687** ± 0.070	**0.755** ± 0.126	0.760 ± 0.080	0.726 ± 0.055	0.956	0.451
Size-universal Ensemble	M8	+	+	+	+				0.609 ± 0.092	0.776 ± 0.054	0.775 ± 0.091	0.714 ± 0.069	0.650	0.219
	M9	+	+	+	+	+			0.675 ± 0.073	**0.815** ± 0.065	0.807 ± 0.073	0.748 ± 0.044	**0.028**	0.859
	M10	+		+	+				0.625 ± 0.048	0.767 ± 0.054	0.785 ± 0.105	0.717 ± 0.053	0.499	0.399
	M11	+		+	+	+			**0.688** ± 0.069	0.753 ± 0.069	**0.826** ± 0.051	0.757 ± 0.049	**<0.01**	0.740
	M12	+		+					0.629 ± 0.058	0.769 ± 0.050	0.782 ± 0.080	0.719 ± 0.054	0.482	0.384
	M13	+		+		+			0.684 ± 0.072	**0.815** ± 0.050	0.805 ± 0.048	**0.756** ± 0.043	**<0.01**	0.788
	M14			+		+			0.678 ± 0.054	0.813 ± 0.074	0.816 ± 0.070	0.747 ± 0.048	**<0.01**	0.818
Size-binned Ensemble	M15	+	+	+	+		+	+	0.678 ± 0.058	0.780 ± 0.072	0.750 ± 0.089	0.735 ± 0.041	0.816	0.644
	M16	+	+	+	+	+	+	+	0.708 ± 0.067	0.818 ± 0.101	0.789 ± 0.065	0.765 ± 0.048	0.078	0.359
	M17	+		+	+		+	+	0.683 ± 0.067	0.787 ± 0.087	0.766 ± 0.066	0.742 ± 0.040	0.688	0.627
	M18	+		+	+	+	+	+	**0.726** ± 0.067	0.822 ± 0.106	0.799 ± 0.057	0.775 ± 0.048	**0.040**	0.144
	M19	+		+			+	+	0.685 ± 0.059	0.786 ± 0.088	0.764 ± 0.071	0.741 ± 0.041	0.679	0.533
	M20	+		+		+	+	+	0.723 ± 0.067	**0.823** ± 0.100	0.797 ± 0.060	0.774 ± 0.048	**0.026**	0.153
	M21			+			+	+	0.676 ± 0.058	0.776 ± 0.103	0.778 ± 0.061	0.746 ± 0.048	0.571	0.599
	M22			+		+	+	+	0.718 ± 0.082	0.816 ± 0.107	0.822 ± 0.048	0.778 ± 0.063	**0.030**	**0.032**
	M23						+	+	0.687 ± 0.070	0.771 ± 0.105	0.778 ± 0.059	0.748 ± 0.053	0.572	0.622
	M24					+	+	+	0.714 ± 0.076	0.818 ± 0.111	0.825 ± 0.05	0.778 ± 0.064	**0.015**	**0.017**
	M25					+	+		0.713 ± 0.077	0.815 ± 0.11	**0.830** ± 0.054	**0.780** ± 0.063	**0.013**	**<0.01**

For the selection of the second-stage classifier, 9 different classifiers including K nearest neighbors [165], LOG, linear support vector machine [170], SVMp, SVMr, decision tree, random forest [171], AdaBoosted Tree [172], and Gaussian Naive Bayes [173], are explored, with hyper-parameters in each classifier tuned using the validation set. For each ensemble model, the classifier which achieved the best performance (averaged over 5 folds) in the validation set is selected as the second-stage classifier. The training and evaluation of the in the validation set is selected as the second-stage classifier. The training and evaluation of the classifier ensembles are implemented using Scikit-learn python package [174].

5.3.4 Evaluation

To investigate the benefits of employing 3D CNN compared to 2D CNN, six additional 2D architectures, CNN4 – CNN9, as described in Table 5.1, were also evaluated. CNN9 is the comparable 2D network architecture of CNN1. CNN5 and CNN7 are simplified version of VGG network [135], which is one of the classic architectures used for the task of natural image classification. These three 2D networks map each 3D volumetric image into a 3-channel 2D image by assigning 3 orthogonal views (axial, coronal and sagittal views centered at the nodule) to different input channels following the approach used in [139, 142, 143]. CNN4, CNN6, and CNN8 replicate the model presented by Shen et al in [146], which utilize 3D input volumes and consider the 3rd dimension (z-axis) as the input channels. The same dataset (including the same image pre-processing, data augmentation and cross validation split) and the training strategy as described in the previous sections were used for the 2D model. The hyper-parameters, including the initial learning rate η_0 and weight decay C parameter, were re-tuned for the 2D model based on the performance tested in the validation set.

The receiver operating characteristic (ROC) curve averaged over 5 cross-validation folds for each classifier is plotted and the respective area under the curve (AUC) and standard deviation (σ) is reported. As a measure of the difference between a pair of ROC curves, the statistical significance p-value (with significance level of 0.05) of the difference is computed based on the DeLong test [175]. The p-values for 5 cross-validation folds are combined using Fisher's method [176, 177]. The statistical tests on ROC curves are implemented using pROC R package [178].

5.4 Results

The performance comparison for all 9 presented CNN models, including 3 3D CNNs and 6 2D CNNs are summarized in Table 5.1. The detailed performance comparison of the best two 3D CNN models, CNN1 and CNN2 are summarized in Table 5.4. CNN1 outperforms all of the other CNNs, thus it is used in the construction of ensemble models listed in Table 5.3. The performance for 7 single models and 18 ensemble models are summarized in Table 5.3. The columns marked by + indicate the classifiers included in the respective model on each row. The performance for each model is summarized in 6 rightmost columns, including overall AUC $\pm$ σ, AUC $\pm$ σ for each size bin, the p-values for the ROC difference compared to M3 (SVMp) and the p-values for the ROC difference compared to M6 (2-bin). M3 and M6 are selected as reference because they are the best traditional size-universal single model and the best traditional size-binned model respectively.

Table 5.4. The summary of performance of best two 3D CNN models, CNN1 and CNN2. The overall AUC and AUC for each size bin are averaged over 5 folds with corresponding standard deviation (σ) reported below.

CNN models	AUC for B6 ± σ	AUC for B8 ± σ	AUC for B12 ± σ	Overall AUC ± σ
CNN1 (CNN-16-2conv)	**0.734** ± 0.051	**0.804** ± 0.089	0.682 ± 0.119	**0.732** ± 0.052
CNN2 (CNN-32-3conv)	0.649 ± 0.058	0.738 ± 0.108	**0.761** ± 0.082	0.698 ± 0.047

The comparison of the ROC curves of 7 single models (M1- M7) is shown in Figure 5.5. Size-universal models (M1-M5) are plotted in solid line and size-binned models (M6 – M7) are plotted in dashed line. The comparison of ROC curves of 4 size-universal ensemble models (M8, M9, M12 and M13) is shown in Figure 5.6. Two single models M3 SVMp and M5 CNN are also shown for reference, since they are the best traditional size-universal single model and the best size-universal single model respectively. Models (M5, M9 and M13) that include CNN are plotted in solid lines and the models (M3, M8 and M12) that are built only with traditional models are plotted in dashed lines. The comparison of ROC curves of 3 size-binned ensemble models (M15, M16 and M25) is shown in Figure 5.7. Two single models M6 2-Bin and M5 CNN (size-universal) are also shown for reference, since they the best traditional single model and the best size-universal model respectively. Models (M5, M16 and M25) that include CNN are plotted in solid lines and the models (M6 and M15) that are built only with traditional models are plotted in dashed lines. For the clarity of the figures, not all ensemble models in Table 5.3 are included in the plots.

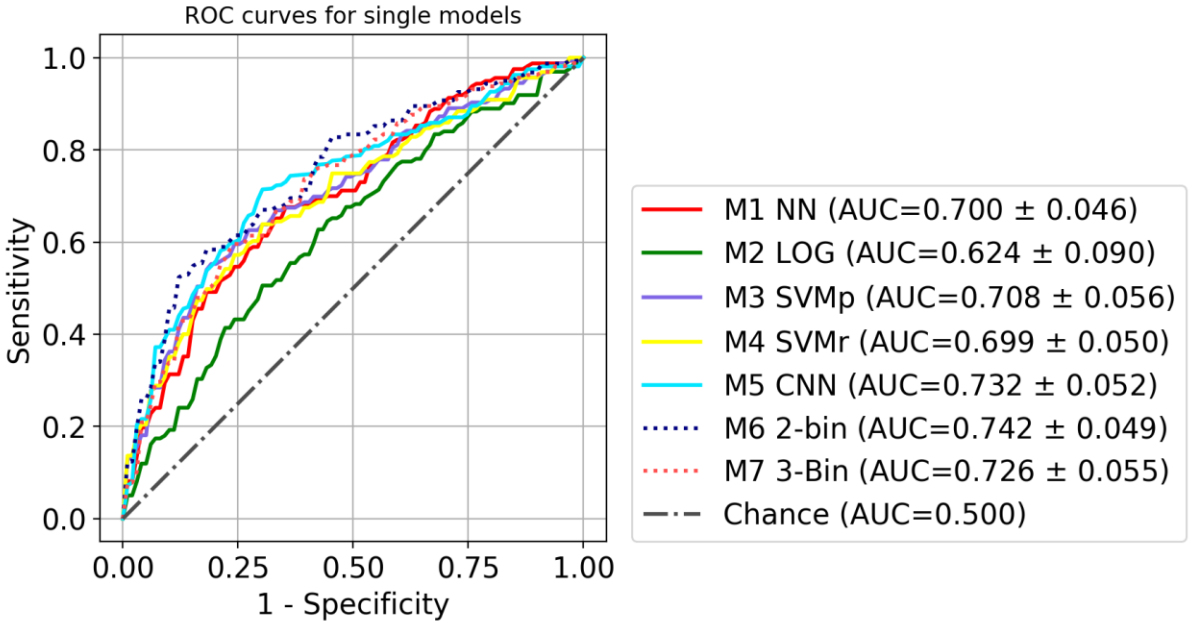

Figure 5.5. The comparison of ROC of single models M1 to M7. The size-universal models (M1 to M5) are plotted in solid lines, and the size-binned models are plotted in dashed lines (M6 to M7).

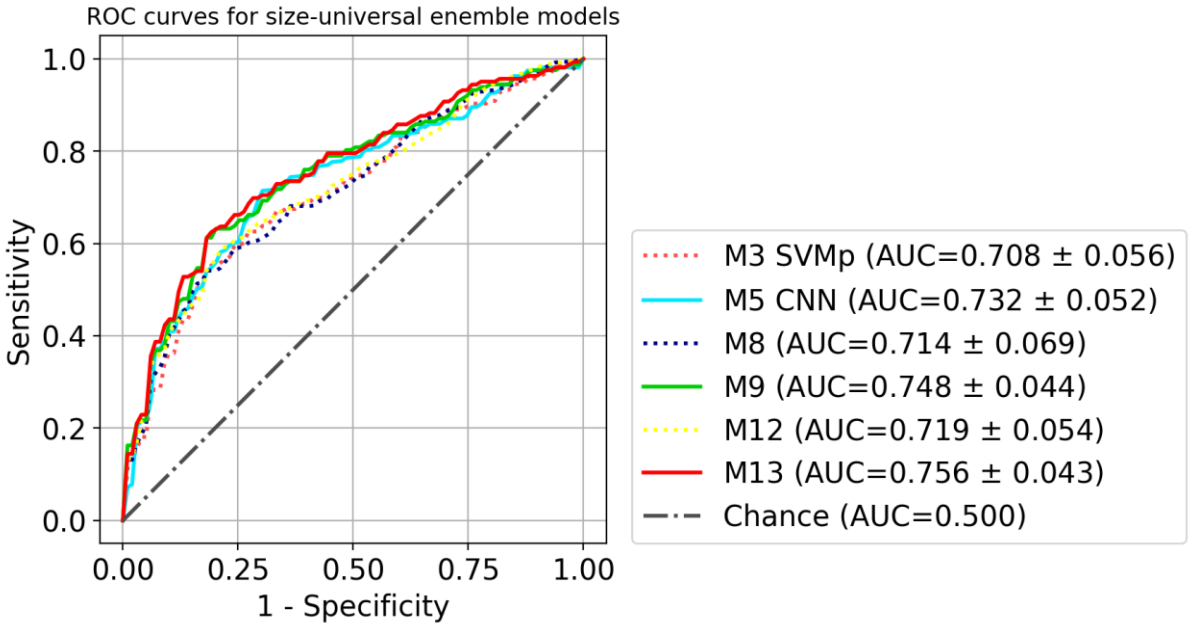

Figure 5.6. The comparison of ROC of size-universal ensemble models M8, M9, M12, and M13. Two single models M3 SVMp and M5 CNN are also shown for reference. Models that include CNN are plotted in solid lines, and the models that are built only with traditional models are plotted in dashed lines.

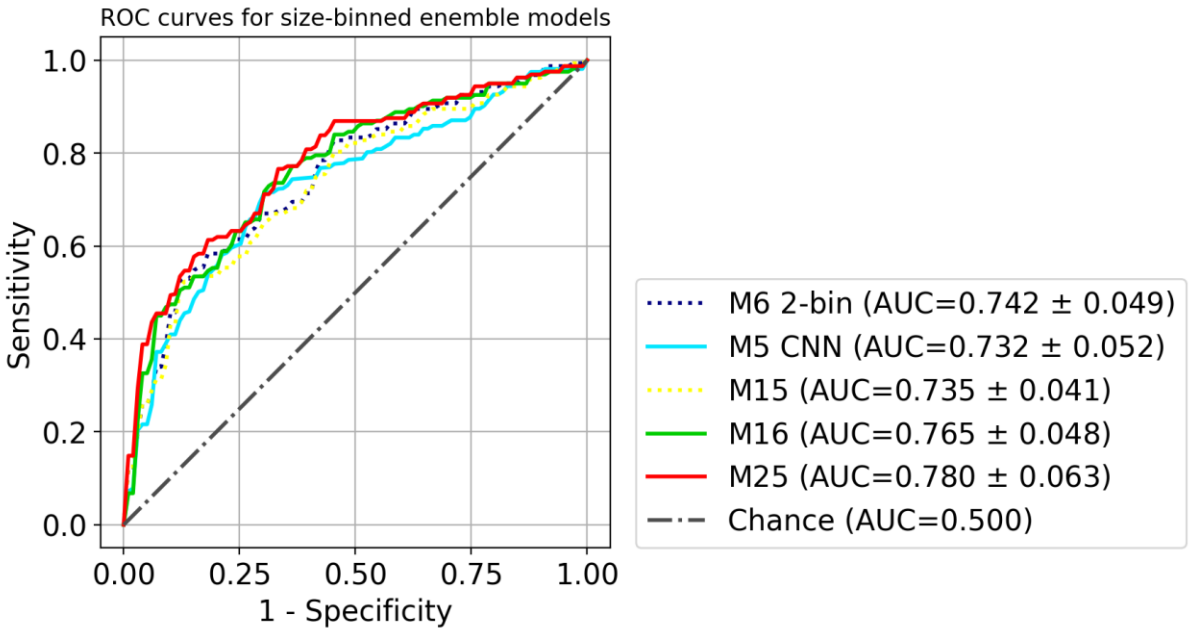

Figure 5.7. The comparison of ROC of size-binned ensemble models M15, M16, and M25. Two single models M6 2-bin and M5 CNN (size-universal) are also shown for reference. Models that include CNN are plotted in solid lines, and the models that are built only with traditional models are plotted in dashed lines.

5.5 Discussion

The task of 3D nodule analysis in lung cancer screening differs from the traditional tasks in 2D imaging in that there is a very wide range of nodule sizes, resulting in the number of pixels on target for a nodule varying by a factor of 57 in our dataset. The difference in resolution is illustrated by Figure 5.8 (g, n). With the CNN1 (16x16x16) model, the largest 27.6% of the nodules were slightly under-sampled with most of these nodules in the B12 group, while there was no under-sampling for the CNN2 (32x32x32) model. This may account for the difference in performance between the two models as shown in Table 5.4. The CNN1 model had better performance for the smaller nodules in B6 and B8 group, while the CNN2 model exhibited better performance for the larger and more detailed nodules in B12 group. A better classification model can be constructed without any re-training by simply

using CNN1 for nodules in B6 and B8 and using CNN2 for nodules in B12, and achieves overall AUC ± σ as 0.761 ± 0.084, although the ROC difference is not statistically significant with respect to CNN1.

The size-universal single classification model of 3D CNN (M5) has been shown to achieve better AUC compared to the size-universal single models (M1- M4) constructed using handcrafted features and traditional machine learning approaches as shown in Figure 5.5 and Table 5.3, although the ROC difference between the 3D CNN model and the best traditional model M3 is not statistically significant (p-value 0.256). Since the exact same 5-fold training and testing partition and evaluation scheme were used, the direct performance comparison demonstrates the strength of the 3D CNN approach with the benefits of eliminating manual feature design and selection, which relies on task-specific expert knowledge and can be rather time consuming. Moreover, due to the much smaller scale of available training examples compared to computer vision applications in natural images [131, 134, 135, 136], it is reasonable to hypothesize that better performance can be obtained by the 3D CNN model, if more training examples are available and thus deeper network architectures can be utilized, based on the relation between the performance and dataset size observed in other studies [146, 179].

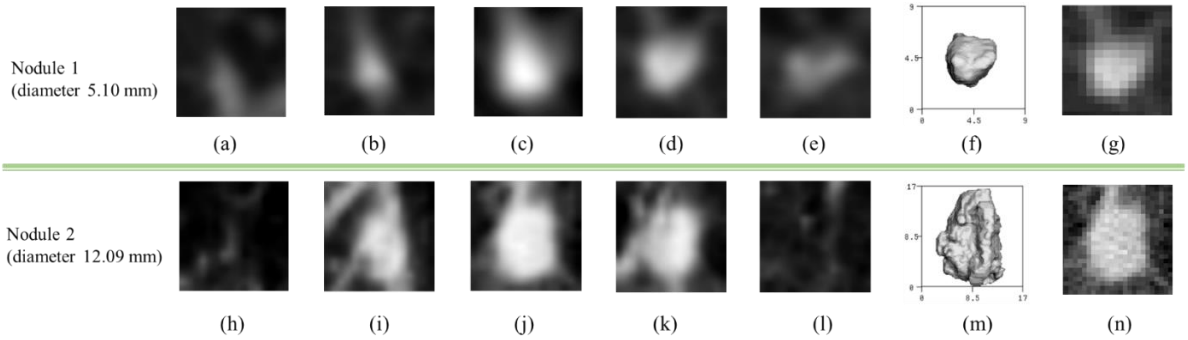

Figure 5.8. (a-e) Five axial CT slices sampled from the input 3D volume of a nodule (nodule 1) with diameter of 5.1 mm. (f) the segmentation of nodule 1 shown in 3D axial view. (h - l) Five axial CT slices sampled from the input 3D volume of a nodule (nodule 2) with diameter of 12.09 mm. (m) the segmentation of nodule 2 shown in 3D axial view. (g) and (n) are axial slices cropped from original CT before rescaling, as the counterpart of (c) and (j) respectively. Since the nodules are first cropped based on nodule size and then scaled to the same image size, the image appearance of the nodule and nearby structures (such as vessels) can be rather different.

The size-binned single models (M6 and M7) outperform all the size-universal single models (M1 – M5) as shown in Figure 5.5 and Table 5.3, although the ROC difference between the best size-binned model M6 and the best size-universal model 3D CNN M5 is not statistically significant (p-value 0.444). It suggests that given more training examples, a size-binned 3D CNN model may potentially achieve better performance than its size-universal counterpart, because of the advantage of considering the nodules of different size range separately. Additionally, in the presented size-universal 3D CNN model, to ensure a uniform target object scale that is usually considered as helpful for the training of CNNs, nodules of different sizes are scaled to image volumes of the same size in pixels. This results in very different image representations of nodules and nearby structures, such as more blurring effect and larger scale of the attached vessels for small nodules, as illustrated in Figure 5.8, and thus has a potential negative effect on the final classification performance. Unfortunately, due to the limited size of current dataset, a size-binned 3D CNN cannot be trained to converge, as it

means only 44.79% of the training set can be used to train a 3D CNN for size bin B6, 28.22% training set for bin size B8, and 26.99% training set for bin size B12. Finally, for all models shown in Table 5.3, the best overall AUC is 0.735 for B6, 0.823 for B8 and 0.830 for B12, which is consistent with classification of small nodules being more challenging [126, 128] due to the larger number of on target pixels for larger nodule.

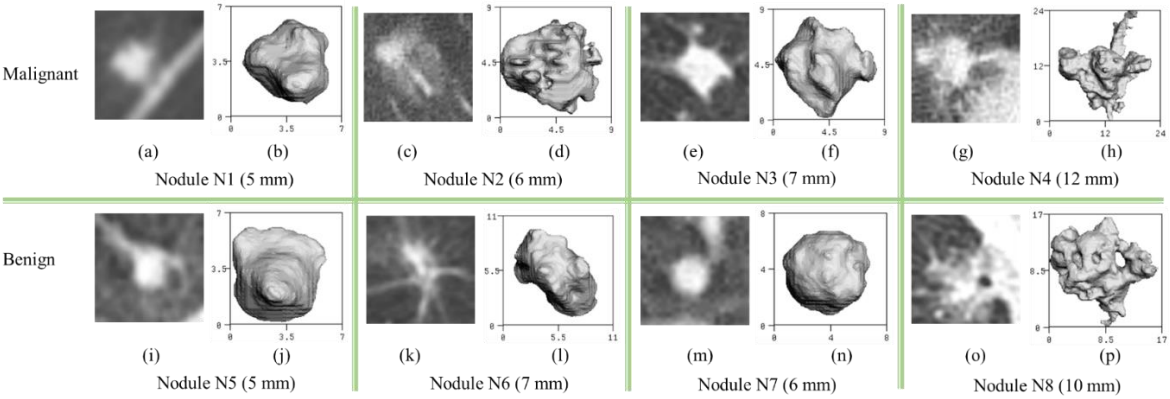

Figure 5.9. Examples of malignant nodules N1 – N4 (a – h) shown on the first row and benign nodules N5 – N8 (i – p) shown on the second row. For each nodule, the central axial slice from CT scan is shown on the left and the segmentation of each nodule is shown in 3D axial view on the right.

The incorporation of 3D CNN model into the ensembles of other traditional models always leads to performance enhancement compared to the ensemble counterparts without the 3D CNN as shown in Table 5.3, Figure 5.6 and Figure 5.7 (solid lines vs. dashed lines). The models with best performance in each of the three categories shown in Table 5.3, including M5 for size-universal single model, M13 for size-universal ensemble model and M25 for size-binned ensemble model, all include 3D CNN in its composition, whereas the simple combination of traditional models, such as M8, M10, M12, M15, M17, M19, M21 and M23, can only lead to negligible performance improvement compared to the best single model, with no statistical significance, as indicated in Table 5.3. The ROC differences between the best

performance ensemble models and the respective best performance traditional single model, i.e., M13 vs. M3, and M25 vs. M6, are statistically significant (p-values < 0.01).

The results on the ensemble models demonstrate that the diversity among the individual models in the composition is the key to the success of the ensembles, which is consistent with the discussion by Kittler et al. [163] and Kuncheva et al. [164]. As illustrated in the examples given in Figure 5.9 and Figure 5.10, misclassifications by different single classifiers (M1- M7) often do not overlap; consequently, different single classifiers usually exhibit complementary advantages in recognizing different image patterns, leading to an optimized classifier ensemble. Since all the traditional models explored in this paper are built upon the same set of handcrafted image features and trained with the same training data, the diversity among them are limited. On the contrary, the 3D CNN model takes the raw image volumes as inputs and learns the features automatically by the network itself, which often provides complementary information about the image patterns to be classified compared to the traditional models [139, 143], and thus can potentially be harnessed to improve the overall ensemble performance.

Excluding single models with inferior performance from the ensemble models can also be beneficial to the ensemble performance as can been seen in the comparison of M13 vs. M9, and M25 vs. M16 in Table 5.3. On the other hand, excluding more single classifiers with inferior performance may be detrimental to the ensemble performance due to the decreased diversity as can been seen in the comparison of M14 vs. M13 in Table 3. Therefore, the number of single models to be included needs to be optimized in the construction of ensemble models.

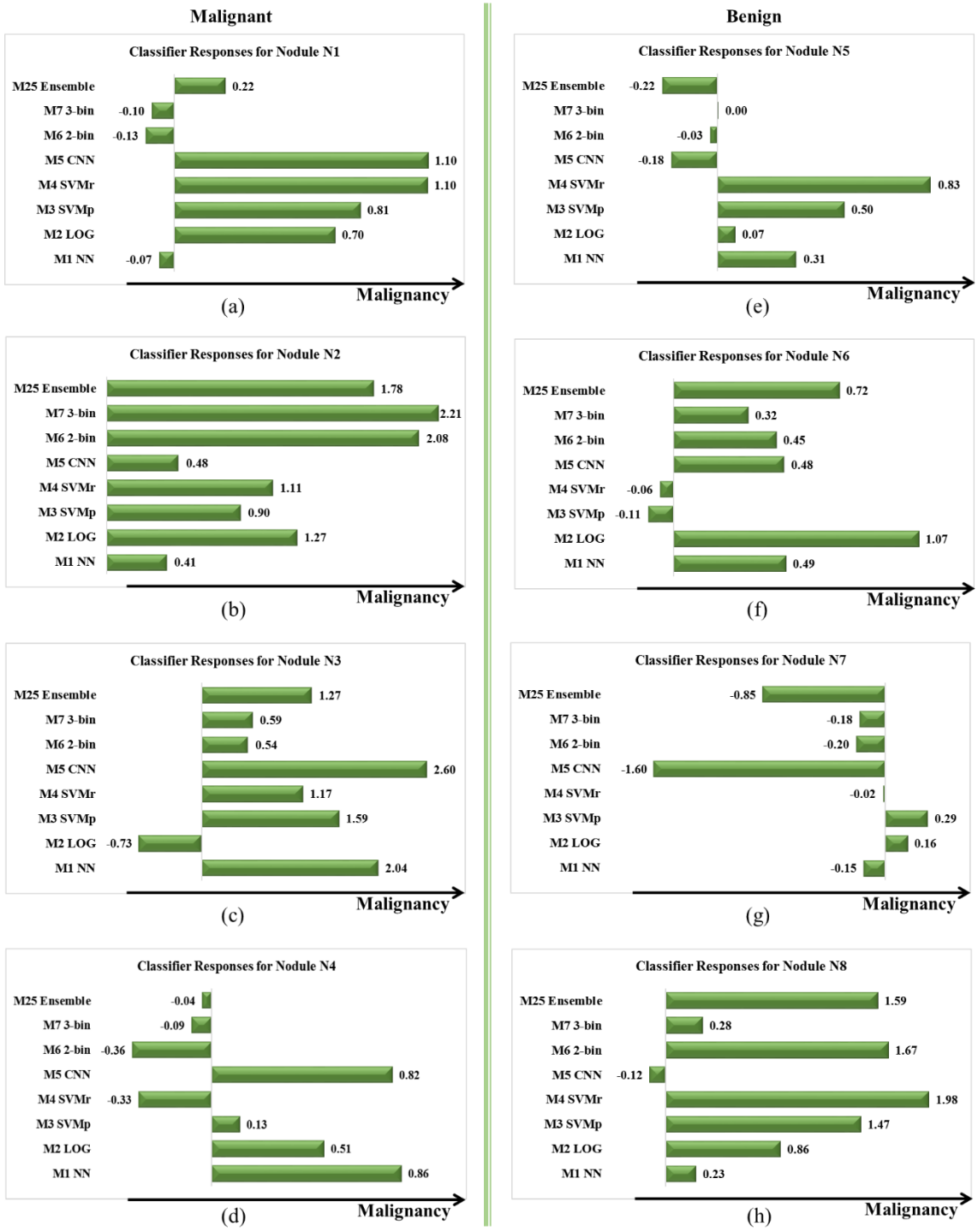

Figure 5.10. The model response output by 7 single models (M1 – M7) and the best ensemble model (M25) for nodules N1 – N8 as defined in Figure 5.9. Malignant nodules N1 – N4 (a – d) are shown in the left column and benign nodules N5 – N8 (e – h) are shown in the right column. For the purpose of direct comparison, the classifier response output by each model has been normalized to zero mean and unit standard deviation. A larger value of response indicates a higher probability for malignancy predicted by the model.

This paper is the first study to employ a 3D CNN trained from scratch to address pulmonary nodule classification from low-dose chest CT. Due to the consideration of the full 3D image volume, it has the potential for better performance [148, 150, 151] compared to other work based on 2D CNN [143, 153]. In addition, unlike using a pre-trained CNN in the study [143] by Buty et al., training a CNN from scratch eliminates the constraint on using the same network architecture as the pre-trained CNN, which may be suboptimal for the specific task of interest [140, 141]. However, in the situation where the computation resource is limited and the training data is insufficient to train a deep neural network, which is often true for automated applications in medical imaging [140], the use of 2D CNN becomes attractive, because it avoids the need of training from scratch and allows the utilization of off-the-shelf deep features and fine-tuning from pre-trained networks that were trained with large-scale annotated natural image datasets [132, 140].

To quantify the benefit of using a 3D CNN architecture for pulmonary nodule classification, six 2D CNN models, CNN4 – CNN9, were also evaluated as shown in Table 5.1. 3D model CNN1 outperforms all the 2D CNN models, suggesting the strength of considering all 3D information and employing of 3D architectures.

The malignancy status of nodules was established in this paper by either biopsy or histology of resected tissue, which should be considered as a much more reliable truth reference compared to the subjective malignancy ratings of radiologists. The nodules exhibit a significant intra-class variation in image appearance on CT scans, i.e., a wide variation in among the same class (benign or malignant), as demonstrated by the examples listed on the same row in Figure 5.9; whereas the inter-class variations can be small, i.e., the appearance of the benign and malignant nodules can be rather similar, which makes the visual

discrimination between benign and malignant nodules challenging, as demonstrated by the examples listed in the same column in Figure 5.9. As shown in the observer study in LUNGx Challenge [122], the inter-observer variations among experienced thoracic radiologists in the task of malignancy rating even on diagnostic chest CT are significant, with AUC ranging from 0.70 to 0.85. As a result, the presented study cannot be directly compared to the studies using subjective malignancy ratings as truth reference [125, 143, 145, 146].

In the future work, there are two possible extensions to consider. First, different ensemble classifiers can be trained for nodules of different size ranges to take advantage of the fact that the performance for some single classifier is superior for one size bin but inferior for another size bin. Second, the complementary information learned by conventional classifiers can be incorporated into the CNN by feeding predictions of conventional models or handcrafted features as input to the FC layers, and then trained from end-to-end to replace the use of ensemble classifiers.

5.6 Conclusion

This study presents a 3D CNN trained from scratch for the challenging task of classifying pulmonary nodule malignancy from low-dose chest CT obtained from the annual screening of lung cancer. The dataset consisting of 326 nodules is constructed with balanced size and class distribution with the malignancy status pathologically confirmed. The experiments were designed to replicate those in the study [126] by Reeves et al. by using the exact same 5-fold training and testing partition, truth definition and evaluation scheme for the direct performance comparison of the 3D CNN and conventional approaches. The results demonstrate three primary advantages of applying 3D CNN to pulmonary nodule classification. First, both the 3D CNN single model (AUC of 0.732) and the ensemble models

with 3D CNN (AUC of 0.780) outperform the respective counterparts constructed using only traditional machine learning models (AUC of 0.708 for the best single traditional model, and AUC of 0.748 for the best ensemble model constructed without CNN). Second, 3D CNN models eliminate the procedure of manual feature design and selection that are required by the traditional machine learning models and rely heavily on the domain-specific expert knowledge. Third, complementary information of nodules can be learned by the 3D CNN and the conventional models, which together are combined to construct an ensemble model with statistically significant performance improvement (p-value < 0.05) compared to any single traditional model in its composition. Although the current best performance model with AUC of 0.780 is insufficient for direct diagnosis in the clinical practice, the automated prediction outcome may be useful in improving the lung cancer screening follow-up protocol which currently depends mainly upon the nodule size.

CHAPTER 6

CONCLUSION

A fully automated framework has been presented in this dissertation for the measurement and evaluation of quantitative image biomarkers from LDCT scans acquired during the annual lung cancer screening. Quantitative image biomarker measurements from the regions of breasts, bones, airway and lungs have been accomplished respectively and demonstrated to be able to potentially assist and even improve the comprehensive interpretation of medical images.

An anatomy directed approach is applied to the analysis of breast region and to the measurements of breast density for women and of gynecomastia quantification for men. The automated density assessment has been demonstrated to be consistent with the subjective reading of an experienced radiologist in 97 of 100 scans. Therefore, breast density assessment from LDCT can potentially serve as a valuable resource providing useful information with respect to breast cancer risk evaluation for many women who have undergone LDCT but not recent mammograms. The automated gynecomastia measurements have been demonstrated to achieve promising performance for the gynecomastia diagnosis with the AUC of 0.86 for the ROC curve and have statistically significant Spearman correlation $r=0.70$ ($p < 0.001$) with the reference categorical grades established by an experienced radiologist. The encouraging results demonstrate the feasibility of fully automated gynecomastia quantification from LDCT, which may assist the early detection as well as the treatment of both gynecomastia and the underlying medical problems, if any, that cause gynecomastia.

Fully automated BMD assessment based on CT image attenuation (HU) is achieved by building upon the model-based segmentation and anatomical labeling of individual vertebral body. Statistically significant (p-value < 0.001) strong correlation can be obtained between BMD_{CT} and the reference BMD_{DXA} at all vertebral levels (T1 – L2). The highest Pearson correlation of 0.857 is achieved between BMD_{DXA} and the average BMD_{CT} of T9-T11. The encouraging results demonstrate the feasibility of fully automated quantitative BMD assessment and the potential of opportunistic osteoporosis screening with concurrent lung cancer screening using LDCT.

A fully automated knowledge-based approach is applied to the segmentation and anatomical labeling of each airway bronchus, which enables the measurements of precise and reproducible airway dimension derived biomarkers, the lumen diameter and wall thickness, for each labeled bronchus. The airway biomarker measurements are evaluated with a longitudinal dataset of 504 LDCT cases, which demonstrates good reproducibility and therefore provides valuable information to aid the diagnosis and treatment of COPD.

For the classification of pulmonary nodule malignancy, a 3D CNN is trained from scratch and demonstrates various advantages over both the traditional machine learning approaches (with the use of hand-crafted 3D image features) and the 2D CNN models. Classifier ensembles of the combinations of the 3D CNN and traditional machine learning models achieve the best performance by taking advantage of the complementary characteristics of the traditional models and the CNN models. Although the current best performance model with AUC of 0.780 is insufficient for direct diagnosis in the clinical practice, the output prediction of the automated system may be useful in assisting radiologist's decision making on the lung cancer screening follow-up plan.

In conclusion, with the recent large-scale implementation of annual lung cancer screening in the US using LDCT, great potential emerges for the concurrent extraction of quantitative image biomarkers from different regions in the chest, including lungs, heart, vertebrae, breasts, etc., which are covered in LDCT scans acquired during annual lung cancer screening. This dissertation has demonstrated the feasibility of fully automated measurement and evaluation of a rich set of quantitative image biomarkers, and the opportunity to significantly enhance the impact of LDCT acquired in the annual lung cancer screening by offering a comprehensive health assessment to each screening participant with no additional imaging or radiation exposure.

REFERENCES

[1] D. Sullivan, N. Obuchowski, L. Kessler, D. Raunig, C. Gatsonis, E. Huang, M. Kondratovich, L. McShane, A. Reeves, D. Barboriak and A. Guimaraes, "Metrology standards for quantitative imaging biomarkers," *Radiology,* vol. 277, no. 3, pp. 813-825, 2015.

[2] FDA–NIH Biomarker Working Group, "BEST (Biomarkers, EndpointS, and other Tools) resource.," [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK326791. [Accessed 1 August 2016].

[3] J. O'connor, E. Aboagye, J. Adams, H. Aerts, S. Barrington, A. Beer, R. Boellaard, S. Bohndiek, M. Brady, G. Brown and D. Buckley, "Imaging biomarker roadmap for cancer studies," *Nature reviews Clinical oncology,* vol. 14, no. 3, p. 169, 2017.

[4] A. Reeves, Y. Xie and S. Liu, "Large-scale image region documentation for fully automated image biomarker algorithm development and evaluation," *Journal of Medical Imaging,* vol. 4, no. 2, p. 024505, 2017.

[5] American Cancer Society, "Key statistics for lung cancer," [Online]. Available: https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html. [Accessed 11 April 2018].

[6] N. Howlader, A. Noone, M. Krapcho, D. Miller, K. Bishop, S. Altekruse, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich and A. Mariotto, SEER Cancer Statistics Review, 1975-2013, Bethesda, MD: National Cancer Institute, 2016.

[7] International Early Lung Cancer Action Program Investigators, "Survival of patients with stage I lung cancer detected on CT screening," *N Engl J Med,* vol. 2006, no. 355, pp. 1763-1771, 2006.

[8] National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *N Engl J Med,* vol. 2011, no. 365, pp. 395-409, 2011.

[9] Centers for Medicare & Medicaid Services, "Decision Memo for Screening for Lung Cancer with Low Dose Computed Tomography (LDCT) (CAG-00439N)," [Online]. Available: http://www.cms.gov/medicare-eoverage-data-base/details/nca-decision-memo.aspx. [Accessed 22 September 2017].

[10] J. Ma, E. M. Ward, R. Smith and A. Jemal, "Annual number of lung cancer deaths potentially avertable by screening in the United States," *Cancer,* vol. 119, no. 7, pp. 1381-1385, 2013.

[11] F. Larke, R. Kruger, C. Cagnon, M. Flynn, M. McNitt-Gray, X. Wu, P. Judy and D. Cody, "Estimated radiation dose associated with low-dose chest CT of average-size participants in the National Lung Screening Trial," *American Journal of Roentgenology,* vol. 197, no. 5, pp. 1165-1169, 2011.

[12] J. Boone, J. Brink, S. Edyvean, W. Huda, W. Leitz, C. McCollough, M. McNitt-Gray, P. Dawson, P. Deluca, S. Seltzer and J. Brunberg, "Radiation dose and image-quality assessment in computed tomography.," *Journal of the ICRU,* vol. 12, no. 1, pp. 9-149, 2012.

[13] C. DeSantis, J. Ma, L. Bryan and A. Jemal, "Breast cancer statistics, 2013," *CA: a cancer journal for clinicians*, vol. 64, no. 1, pp. 52-62, 2014.

[14] K. M. Ray, E. R. Price and B. N. Joe, "Breast density legislation: mandatory disclosure to patients, alternative screening, billing, reimbursement," *Am J Roentgenol*, vol. 204, no. 2, pp. 257-260, 2015.

[15] H. Carlson, "Approach to the patient with gynecomastia," *J. Clin. Endocrinol. Metab*, vol. 96, no. 1, pp. 15-21, 2011.

[16] F. Nuttall, "Gynecomastia as a physical finding in normal men," *J. Clin. Endocrinol. Metab*, vol. 48, no. 2, pp. 338-340, 1979.

[17] R. Cooper, B. Gunter and L. Ramamurthy, " Mammography in men," *Radiology*, vol. 191, no. 3, pp. 651-656, 1994.

[18] N. Wright, A. Looker, K. Saag, J. Curtis, E. Delzell and S. Randall, "The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine," *Journal of Bone and Mineral Research*, vol. 29, no. 11, pp. 2520-2526, 2014.

[19] World Health Organization, Prevention and management of osteoporosis: report of a WHO scientific group (No. 921), Diamond Pocket Books (P) Ltd, 2003.

[20] J. Reginster and N. Burlet, "Osteoporosis: a still increasing prevalence," *Bone*, vol. 38, no. 2, pp. 4-9, 2006.

[21] G. Sturton, C. Persson and P. Barnes, "Small airways: an important but neglected target in the treatment of obstructive airway diseases," *Trends in pharmacological sciences*, vol. 29, no. 7, pp. 340-345, 2008.

[22] E. Arias and B. Smith, "Deaths: preliminary data for 2001," *Natl. Vital Stat.*, vol. 51, no. 5, p. 1–44, 2003.

[23] A. Reeves, A. Chan, D. Yankelevitz, C. Henschke, B. Kressler and W. Kostis, "On measuring the change in size of pulmonary nodules," *IEEE transactions on medical imaging*, vol. 25, no. 4, pp. 435-450, 2006.

[24] J. Chen, S. Chan, N. Lu, Y. Li, Y. Tsai, P. Huang, C. Chang and M. Su, "Opportunistic breast density assessment in women receiving low-dose chest computed tomography screening," *Academic radiology*, vol. 23, no. 9, pp. 1154-1161, 2016 .

[25] W. K. Moon, C. M. Lo, J. M. Goo, M. S. Bae, J. M. Chang and C. S. Huang, "Quantitative analysis for breast density estimation in low dose chest CT scans," *J Med Syst*, vol. 38, no. 3, pp. 1-9, 2014.

[26] M. Salvatore, L. Margolies, M. Kale, J. Wisnivesky, S. Kotkin, C. I. Henschke and D. F. Yankelevitz, " Breast density: comparison of chest CT with mammography.," *Radiology*, vol. 270, no. 1, pp. 67-73, 2014.

[27] M. Lin, J. H. Chen, X. Wang, S. Chan, S. Chen and M. Y. Su, "Template-based automatic breast segmentation on MRI by excluding the chest region," *Med Phys*, vol. 40, no. 12, pp. 122-301, 2013.

[28] E. A. Sickles, C. J. D'Orsi, L. W. Bassett, C. M. Appleton, W. A. Berg and E. S. Burnside, ACR BI-RADS® Atlas Breast Imaging Reporting and Data System, Reston, VA: American College of Radiology, 2013.

[29] E. B. Sonnenblick, L. R. Margolies, J. R. Szabo, L. M. Jacobs, N. Patel and K. A. Lee, "Digital breast tomosynthesis of gynecomastia and associated findings—a pictorial review.," *Clinical imaging,* vol. 38, no. 5, pp. 565-570, 2014.

[30] R. J. Hooley, K. L. Greenberg, R. M. Stackhouse, J. L. Geisel and R. S. Butler, "Screening US in patients with mammographically dense breasts: initial experience with Connecticut Public Act 09-41," *Radiology,* vol. 265, no. 1, pp. 59-69, 2012.

[31] H. Carlson, "Gynecomastia," *N Engl J Med,* vol. 303, no. 14, pp. 795-799, 1980 .

[32] A. Eckman and A. Dobs, "Drug-induced gynecomastia. Expert opinion on drug safety," vol. 7, no. 6, pp. 691-702, 2008.

[33] G. Braunstein, "Gynecomastia," *N Engl J Med,* vol. 357, no. 12, pp. 1229-1237, 2007.

[34] E. Sonnenblick, M. Salvatore, J. Szabo, K. Lee and L. Margolies, "Incremental role of mammography in the evaluation of gynecomastia in men who have undergone chest CT," *Am J Roentgenol,* vol. 207, no. 2, pp. 234-240, 2016.

[35] M. Madhukar and A. Chetlen, "Multimodality imaging of benign and malignant male breast disease," *Clin Radiol,* vol. 68, no. 12, pp. e698-e706, 2013.

[36] E. Klang, N. Rozendorn, S. Raskin, O. Portnoy, M. Sklair, E. Marom, E. Konen and M. Amitai, "CT measurement of breast glandular tissue and its association with testicular cancer," *European radiology,* vol. 27, no. 2, pp. 536-542, 2017.

[37] P. Casti, A. Mencattini, M. Salmeri and A. Ancona, "Automatic detection of the nipple in screen-film and full-field digital mammograms using a novel Hessian-based method," *Journal of digital imaging,* vol. 26, no. 5, pp. 948-957, 2013.

[38] R. Chandrasekhar and Y. Attikiouzel, "A simple method for automatically locating the nipple on mammograms," *IEEE transactions on medical imaging,,* vol. 16, no. 5, pp. 483-494, 1997.

[39] J. Chakraborty, A. Midya, Mukhopadhyay and R. S., "Detection of the nipple in mammograms with Gabor filters and the Radon transform," *Biomedical Signal Processing and Control,* vol. 15, pp. 80-89, 2015.

[40] M. Jas, S. Mukhopadhyay, J. Chakraborty, A. Sadhu and N. Khandelwal, "A heuristic approach to automated nipple detection in digital mammograms," *Journal of digital imaging,* vol. 26, no. 5, pp. 932-940, 2013.

[41] F. Yin, M. Giger, C. Vyborny and R. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," *Medical Physics,* vol. 21, no. 3, pp. 445-452, 1994.

[42] S. Yang, H. Hsu, G. Hsu and P. Chung, "3D localization of clustered microcalcifications using cranio-caudal and medio-lateral oblique views," *Computerized Medical Imaging and Graphics,* vol. 29, no. 7, pp. 521-532, 2005.

[43] X. Zhou, T. Kano, H. Koyasub and S. Li, "Automated assessment of breast tissue density in non-contrast 3D CT images without image segmentation based on a deep CNN," *In SPIE Medical Imaging,* pp. 101342Q-1, March 2017.

[44] X. Zhou, M. Han, T. Hara and H. Fujita, "Automated segmentation of mammary gland regions in non-contrast X-ray CT images," *Computerized Medical Imaging and Graphics,* vol. 32, no. 8, pp. 699-709, 2008.

[45] S. Liu, M. L.R., Y. Xie, D. Yankelevitz, C. Henschke and A. Reeves, "Fully automated breast density assessment from low-dose chest CT," *In SPIE Medical Imaging,* p. 101340R, March 2017.

[46] S. Liu, M. Salvatore, D. Yankelevitz, C. Henschke and A. Reeves, "Segmentation of the whole breast from low-dose chest CT images," *In SPIE Medical Imaging,* p. 94140I, March 2015.

[47] X. Zhou, T. Kano, Y. Cai, S. Li, X. Zhou and T. Hara, "Automatic quantification of mammary glands on non-contrast X-ray CT by using a novel segmentation approach," *In SPIE Medical Imaging,* p. 97851Z, March 2016.

[48] C. Gwo, A. Gwo, C. Wei and P. Huang, "Identification of breast contour for nipple segmentation in breast magnetic resonance images," *Medical physics,* vol. 41, no. 2, 2014.

[49] J. Padgett, A. M. Biancardi, C. I. Henschke, D. Yankelevitz and A. P. Reeves, "Local noise estimation in low-dose chest CT images," *Int J Comput Assist Radiol Surg,* vol. 9, no. 2, pp. 221-229, 2014.

[50] S. Liu, Y. Xie and A. P. Reeves, "Segmentation of the sternum from low-dose chest CT images," *In SPIE Medical Imaging,* p. 941403, March 2015.

[51] J. Lee and A. P. Reeves, "Segmentation of individual ribs from low-dose chest CT," *In SPIE Medical Imaging,* p. 76243J, March 2010.

[52] S. Liu, E. Sonnenblick, L. Azour, D. Yankelevitz, C. Henschke and A. Reeves, "Fully automated gynecomastia quantification from low-dose chest CT," *In SPIE Medical Imaging,* p. 1057524, February 2018.

[53] V. K. Reed, W. A. Woodward, L. Zhang and E. A. Strom, "Automatic segmentation of whole breast using atlas approach and deformable image registration," *International journal of radiation oncology, biology, physics* , vol. 73, no. 5, pp. 1493-1500, 2009.

[54] Y. Xie, J. Padgett, A. Biancardi and A. Reeves, "Automated aorta segmentation in low-dose chest CT images," *Int J Comput Assist Radiol Surg,* vol. 9, no. 2, pp. 211-219, 2014.

[55] Y. Xie, Y. Htwe, J. Padgett, C. Henschke, D. Yankelevitz and A. Reeves, "Automated aortic calcification detection in low-dose chest CT images," *In SPIE Medical Imaging,* p. 90350P, March 2014.

[56] Y. Xie, M. Cham, C. Henschke, D. Yankelevitz and A. Reeves, "Automated coronary artery calcification detection on low-dose chest CT images," *In SPIE Medical Imaging,* p. 90350F, March 2014.

[57] Y. Zheng, F. Vega-Higuera, S. K. Zhou and D. Comaniciu, "Fast and automatic heart isolation in 3D CT volumes: Optimal shape initialization," *In International Workshop on Machine Learning in Medical Imaging,* pp. 84-91, September 2010.

[58] J. Lee, A. Biancardi, A. Reeves, D. Yankelevitz and C. Henschke, "Estimation of anatomical locations using standard frame of reference in chest CT scans," *In Proc IEEE Eng Med Biol Soc,* p. 5809–5812, September 2009.

[59] M. Betke, H. Hong, D. Thomas and C. Prince, "Landmark detection in the chest and registration of lung surfaces with an application to nodule registration," *Medical Image Analysis,* vol. 7, no. 3, pp. 265-281, 2003.

[60] T. Hayashi, H. Chen, K. Miyamoto, X. Zhou, T. Hara and H. Fujita, "Computer-Aided Image Analysis for Vertebral Anatomy on X-Ray CT Images," *CIMI,* pp. 159-184, 2014.

[61] J. Burns, J. Yao, H. Muñoz and R. Summers, "Automated detection, localization, and classification of traumatic vertebral body fractures in the thoracic and lumbar spine at CT," *Radiology,* vol. 278, no. 1, pp. 64-73, 2015.

[62] T. Hayashi, X. Zhou, H. Chen, T. Hara, K. Miyamoto and T. Kobayashi, "Automated extraction method for the center line of spinal canal and its application to the spinal curvature quantification in torso X-ray CT images," *In SPIE Medical Imaging,* p. 76233F, March 2010.

[63] A. Reeves, S. Liu and Y. Xie, "Image segmentation evaluation for very-large datasets," *In SPIE Medical Imaging,* p. 97853J, March 2016.

[64] X. Zhou, T. Hayashi, M. Han, H. Chen, T. Hara and F. R. Yokoyam, "Automated segmentation and recognition of the bone structure in non-contrast torso CT images using implicit anatomical knowledge," *In SPIE Medical Imaging,* p. 72593S, March 2009.

[65] T. Klinder, C. Lorenz, J. Von Berg, S. Dries, T. Bülow and J. Ostermann, "Automated model-based rib cage segmentation and labeling in CT images," *In Proc. MICCAI,* pp. 195-202, October 2007.

[66] J. Staal, B. van Ginneken and M. Viergever, "Automatic rib segmentation and labeling in computed tomography scans using a general framework for detection, recognition and segmentation of objects in volumetric data," *Med Image Anal,* vol. 11, no. 1, pp. 35-46, 2007.

[67] J. Lee, A. P. Reeves, S. V. Fotin, T. Apanasovich and D. F. Yankelevitz, "Human airway measurement from CT images," *In SPIE Medical Imaging,* p. 691518, March 2008.

[68] A. P. Reeves, A. M. Biancardi, D. F. Yankelevitz, S. Fotin, B. M. Keller, A. Jirapatnakul and J. Lee, "A public image database to support research in computer aided diagnosis," *EMBC,* p. 3715–3718, September 2009.

[69] A. P. Reeves, A. M. Biancardi, T. V. Apanasovich, C. R. Meyer, H. MacMahon and E. J. van Beek, "The lung image database consortium (LIDC) a comparison of different size metrics for pulmonary nodule measurements," *Acad Radiol,* vol. 14, no. 12, pp. 1475-1485, 2007.

[70] G. Guglielmi, S. Muscarella and A. Bazzocchi, "Integrated imaging approach to osteoporosis: state-of-the-art review and update," *Radiographics,* vol. 31, no. 5, pp. 1343-1364, 2011.

[71] R. Burge, B. Dawson-Hughes, D. Solomon, J. Wong, A. King and A. Tosteson, "Incidence and economic burden of osteoporosis-related fractures in the United States, 2005–2025.," *Journal of bone and mineral research,* vol. 22, no. 3, pp. 465-475, 2007.

[72] G. Guglielmi, F. Ferrari and A. Bazzocchi, "Bone mineral density and quantitative imaging," *In Pitfalls in Diagnostic Radiology,* pp. 109-132, January 2015.

[73] E. Alacreu, D. Moratal and E. Arana, "Opportunistic screening for osteoporosis by routine CT in Southern Europe," *Osteoporos Int,* vol. 28, no. 3, pp. 983-990, 2017.

[74] M. Boomsma, I. Slouwerhof, J. van Dalen, M. Edens, D. Mueller, J. Milles and M. Maas, "Use of internal references for assessing CT density measurements of the pelvis as replacement for use of an external phantom," *Skeletal radiology,* vol. 44, no. 11, pp. 1597-1602 , 2015.

[75] Y. Miyabara, D. Holmes III, J. Camp, V. Miller and A. Kearns, "Comparison of calibrated and uncalibrated bone mineral density by CT to DEXA in menopausal women," *Climacteric,* vol. 15, no. 4, pp. 374-381, 2012.

[76] D. Mueller, A. Kutscherenko, H. Bartel, A. Vlassenbroek, P. Ourednicek and J. Erckenbrecht, "Phantom-less QCT BMD system as screening tool for osteoporosis without additional radiation," *Eur J Radiol,* vol. 79, no. 3, pp. 375-381, 2011.

[77] P. Pickhardt, L. Lee, A. Muñoz del Rio, T. Lauder, R. Bruce and R. Summers, "Simultaneous screening for osteoporosis at CT colonography: bone mineral density assessment using MDCT attenuation techniques compared with the DXA reference standard," *J Bone Miner Res,,* vol. 26, no. 9, pp. 2194-2203, 2011.

[78] T. Hayashi, H. Chen, K. Miyamoto, X. Zhou, T. Hara and R. Yokoyama, "Analysis of bone mineral density distribution at trabecular bones in thoracic and lumbar vertebras using X-ray CT images," *J Bone Miner Metab,* vol. 29, no. 2, pp. 174-185, 2011.

[79] W. Tay, C. Chui, S. Ong and A. Ng, "Osteoporosis screening using areal bone mineral density estimation from diagnostic CT images," *Academic radiology,* vol. 19, no. 10, pp. 1273-1282, 2012.

[80] R. Summers, N. Baecher, J. Yao, J. Liu, P. Pickhardt, J. Choi and S. Hill, "Feasibility of simultaneous CT colonography and fully-automated bone mineral densitometry in a single examination," *J Comput Assist Tomogr,* vol. 35, no. 2, p. 212, 2011.

[81] X. Zhou, T. Hayashi, H. Chen, T. Hara, Y. R. and K. M., "Automated measurement of bone-mineral-density (BMD) values of vertebral bones based on X-ray torso CT images," *EMBC,* pp. 3573-3576, September 2009.

[82] C. Buckens, G. Dijkhuis, B. de Keizer, H. Verhaar and P. de Jong, "Opportunistic screening for osteoporosis on routine computed tomography? An external validation study," *European radiology,* vol. 25, no. 7, pp. 2074-2079, 2015.

[83] M. Marinova, B. Edon, K. Wolter, B. Katsimbari, H. Schild and H. Strunk, " Use of routine thoracic and abdominal computed tomography scans for assessing bone mineral density and detecting osteoporosis," *Curr Med Res Opin,* vol. 31, no. 10, pp. 1871-1881, 2015.

[84] J. Schreiber, P. Anderson and W. Hsu, "Use of computed tomography for assessing bone mineral density," *Neurosurgical focus,* vol. 37, no. 1, p. E4, 2014.

[85] S. Lee, C. Chung, S. Oh and S. Park, "Correlation between bone mineral density measured by dual-energy X-ray absorptiometry and Hounsfield units measured by diagnostic CT in lumbar spine," *J Korean Neurosurg Soc,* vol. 54, no. 5, pp. 384-389, 2013.

[86] Y. Kim, J. Kim, S. Yoon, J. Lee, C. Lee, C. Shin and Y. Park, "Vertebral bone attenuation on low-dose chest CT: quantitative volumetric analysis for bone fragility assessment," *Osteoporos Int,* vol. 28, no. 1, pp. 329-338, 2017.

[87] S. Nishihara, H. Fujita, T. Iida, A. Takigawa, T. Hara and X. Zhou, "A preliminary study for an automatic recognition algorithm for the central part of a vertebral body using abdominal X-ray CT images," *Comput Med Imaging Graph,* vol. 29, no. 4, pp. 259-266, 2005.

[88] J. Burns, J. Yao and R. Summers, "Vertebral Body Compression Fractures and Bone Density: Automated Detection and Classification on CT Images," *Radiology,* vol. 284, no. 3, pp. 788-797., 2017.

[89] M. Gruber, J. Bauer, M. Dobritz, A. Beer, P. Wolf and K. Woertler, "Bone mineral density measurements of the proximal femur from routine contrast-enhanced MDCT data sets correlate with dual-energy X-ray absorptiometry," *European radiology,* vol. 23, no. 2, pp. 505-512, 2013.

[90] S. Liu, Y. Xie and A. Reeves, "Automated 3D closed surface segmentation: application to vertebral body segmentation in CT images," *J Comput Assist Radiol Surg,* vol. 11, no. 5, pp. 789-801, 2016.

[91] H. Ritzel, M. Amling, M. Pösl, M. Hahn and G. Delling, "The thickness of human vertebral cortical bone and its changes in aging and osteoporosis: a histomorphometric analysis of the complete spinal column from thirty-seven autopsy specimens," *Journal of Bone and Mineral Research,* vol. 12, no. 1, pp. 89-95, 1997.

[92] S. Tan, J. Yao, M. M. Ward, L. Yao and R. M. Summers, "3D Multi-scale level set segmentation of vertebrae.," *ISBI,* pp. 896-899, April 2007.

[93] Insight Software Consortium, The ITK Software Guide Fourth Edition Updated for ITK, 4.8 ed., Kitware Inc, 2015.

[94] V. Caselles, R. Kimmel and S. G., "Geodesic active contours," *International Journal of Computer Vision,* vol. 22, no. 1, pp. 61-79, 1997.

[95] Y. Xie, S. Liu, A. Miller, J. Miller, S. Markowitz and A. Akhund, "Coronary artery calcification identification and labeling in low-dose chest CT images," *In SPIE Medical Imaging,* p. 101340L, March 2017.

[96] H. Tiddens, P. Pare, J. Hogg, W. Hop, R. Lambert and J. De Jongste, "Cartilaginous airway dimensions and airflow obstruction in human lungs," *American journal of respiratory and critical care medicine,* vol. 152, no. 1, pp. 260-266, 1995.

[97] J. Hogg, F. Chu, S. Utokaparch, R. Woods, W. Elliott, L. Buzatu, R. Cherniack, R. Rogers, F. Sciurba, H. Coxson and P. Paré, "The nature of small-airway obstruction in chronic obstructive pulmonary disease," *New England Journal of Medicine,* vol. 350, no. 26, pp. 2645-2653, 2004.

[98] P. Berger, V. Perot, P. Desbarats, J. Tunon-de-Lara, R. Marthan and F. Laurent, "Airway wall thickness in cigarette smokers: quantitative thin-section CT assessment," *Radiology*, vol. 235, no. 3, pp. 1055-1064, 2005.

[99] X. Xie, P. de Jong, M. Oudkerk, Y. Wang, N. ten Hacken, J. Miao, G. Zhang, G. de Bock and R. Vliegenthart, "Morphological measurements in computed tomography correlate with airflow obstruction in chronic obstructive pulmonary disease: systematic review and meta-analysis," *European radiology*, vol. 22, no. 10, pp. 2085-2093, 2012.

[100] Y. Nakano, S. Muro, H. Sakai, T. Hirai, K. Chin, M. Tsukino, K. Nishimura, H. Itoh, P. Paré, J. Hogg and M. Mishima, "Computed tomographic measurements of airway dimensions and emphysema in smokers: correlation with lung function," *American journal of respiratory and critical care medicine*, vol. 162, no. 3, pp. 1102-1108, 2000.

[101] T. Yamashiro, S. Matsuoka, R. Estépar, M. Dransfield, A. Diaz, J. Reilly, S. Patz, S. Murayama, E. Silverman, H. Hatabu and G. Washko, "Quantitative assessment of bronchial wall attenuation with thin-section CT: an indicator of airflow limitation in chronic obstructive pulmonary disease," *American Journal of Roentgenology*, vol. 195, no. 2, pp. 363-369, 2010.

[102] H. Coxson, B. Quiney, D. Sin, L. Xing, A. McWilliams, J. Mayo and S. Lam, "Airway wall thickness assessed using computed tomography and optical coherence tomography," *American journal of respiratory and critical care medicine*, vol. 177, no. 11, pp. 1201-1206, 2008.

[103] S. Matsuoka, Y. Kurihara, K. Yagihashi, M. Hoshino and Y. Nakajima, "Airway dimensions at inspiratory and expiratory multisection CT in chronic obstructive pulmonary disease: correlation with airflow limitation," *Radiology*, vol. 248, no. 3, pp. 1042-1049, 2008.

[104] T. Achenbach, O. Weinheimer, A. Biedermann, S. Schmitt, D. Freudenstein, E. Goutham, R. Kunz, R. Buhl, C. Dueber and C. Heussel, "MDCT assessment of airway wall thickness in COPD patients using a new method: correlations with pulmonary function tests," *European radiology*, vol. 18, no. 12, pp. 2731-2738, 2008.

[105] A. Diaz, F. Rahaghi, J. Ross, R. Harmouche, J. Tschirren, R. Estépar and G. Washko, "Understanding the contribution of native tracheobronchial structure to lung function: CT assessment of airway morphology in never smokers," *Respiratory research*, vol. 16, no. 1, p. 23, 2015.

[106] A. Dijkstra, D. Postma, N. ten Hacken, J. Vonk, M. Oudkerk, v. O. P.M., P. Zanen, F. Hoesein, B. van Ginneken, M. Schmidt and H. Groen, "Low-dose CT measurements of airway dimensions and emphysema associated with airflow limitation in heavy smokers: a cross sectional study," *Respiratory research*, vol. 14, no. 1, p. 11, 2013.

[107] M. Wielpütz, M. Eichinger, O. Weinheimer, S. Ley, M. Mall, M. Wiebel, A. Bischoff, H. Kauczor, C. Heussel and M. Puderbach, "Automatic airway analysis on multidetector computed tomography in cystic fibrosis: correlation with pulmonary function testing," *Journal of thoracic imaging*, vol. 28, no. 2, pp. 104-113, 2013.

[108] J. Petersen, M. Nielsen, P. Lo, L. Nordenmark, J. Pedersen, M. Wille, A. Dirksen and M. de Bruijne, "Optimal surface segmentation using flow lines to quantify airway abnormalities in chronic obstructive pulmonary disease," *Medical image analysis*, vol. 18, no. 3, pp. 531-541, 2014.

[109] P. Paré, T. Nagano and H. Coxson, "Airway imaging in disease: gimmick or useful tool?," *Journal of Applied Physiology,* vol. 113, no. 4, pp. 636-646, 2012.

[110] O. Mets, P. De Jong, B. Van Ginneken, H. Gietema and J. Lammers, "Quantitative computed tomography in COPD: possibilities and limitations," *Lung,* vol. 190, no. 2, pp. 133-145, 2012.

[111] H. Coxson, "Quantitative computed tomography assessment of airway wall dimensions: current status and potential applications for phenotyping chronic obstructive pulmonary disease," *Proceedings of the American Thoracic Society,* vol. 5, no. 9, pp. 940-945, 2008.

[112] J. Hogg, P. Macklem and W. Thurlbeck, "Site and nature of airway obstruction in chronic obstructive lung disease," *New England Journal of Medicine,* vol. 278, no. 25, pp. 1355-1360, 1968.

[113] M. Yana, K. Sekizawa, T. Ohrui, H. Sasaki and T. Takishima, "Site of airway obstruction in pulmonary disease: direct measurement of intrabronchial pressure," *Journal of Applied Physiology,* vol. 72, no. 3, pp. 1016-1023, 1992.

[114] M. Hasegawa, Y. Nasuhara, Y. Onodera, H. Makita, K. Nagai, S. Fuke, Y. Ito, T. Betsuyaku and M. Nishimura, "Airflow limitation and airway dimensions in chronic obstructive pulmonary disease," *American journal of respiratory and critical care ,* vol. 173, no. 12, pp. 1309-1315, 2006.

[115] Y. Nakano, J. Wong, P. de Jong, L. Buzatu, T. Nagao, H. Coxson, W. Elliott, J. Hogg and P. Paré, "The prediction of small airway dimensions using computed tomography," *American journal of respiratory and critical care medicine,* vol. 171, no. 2, pp. 142-146, 2005.

[116] J. Tschirren, E. Hoffman, G. McLennan and M. Sonka, "Intrathoracic airway trees: segmentation and airway morphology analysis from low-dose CT scans," *IEEE transactions on medical imaging,* vol. 24, no. 12, pp. 1529-1539, 2005.

[117] M. Hackx, A. Bankier and P. Gevenois, "Chronic obstructive pulmonary disease: CT quantification of airways disease," *Radiology,* vol. 265, no. 1, pp. 34-48, 2012.

[118] J. Petersen, M. Wille, L. Rakêt, A. Feragen, J. Pedersen, M. Nielsen, A. Dirksen and M. de Bruijne, "Effect of inspiration on airway dimensions measured in maximal inspiration CT images of subjects without airflow limitation," *European radiology,* vol. 24, no. 9, pp. 2319-2325, 2014.

[119] F. Netter, Atlas of human anatomy, Elsevier Health Sciences, 2010.

[120] J. Lee and A. P. Reeves, "Segmentation of the airway tree from chest CT using local volume of interest," *In Proc. of Second International Workshop on Pulmonary Image Analysis,* pp. 273-284, 2009.

[121] V. Rose, M. White, C. Klabunde, M. Nadel, T. Richards, T. McNeel, J. Rodriguez and P. Marcus, "Use of Lung Cancer Screening Tests in the United States: Results from the 2010 National Health Interview Survey," *Cancer epidemiology,* vol. 21, no. 7, pp. 1049-1059, 2012.

[122] S. Armato, K. Drukker, F. Li, L. Hadjiiski, G. Tourassi, R. Engelmann, M. Giger, G. Redmond, K. Farahani, J. Kirby and L. Clarke, "LUNGx Challenge for computerized lung nodule classification," *Journal of Medical Imaging,* vol. 3, no. 4, p. 044506, 2016.

[123] M. Aoyama, Q. Li, S. Katsuragawa, F. Li and S. Sone, "Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images," *Medical Physics,* vol. 30, no. 3, pp. 387-394, 2003.

[124] S. Armato, M. Altman, J. Wilkie, S. Sone, F. Li and A. Roy, "Automated lung nodule classification following automated nodule detection on CT: A serial approach," *Medical Physics,* vol. 30, no. 6, pp. 1188-1197, 2003.

[125] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao and Z. Liang, "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *Journal of digital imaging,* vol. 28, no. 1, pp. 99-115, 2015.

[126] A. Reeves, Y. Xie and A. Jirapatnakul, "Automated pulmonary nodule CT image characterization in lung cancer screening," *International journal of computer assisted radiology and surgery,* vol. 11, no. 1, pp. 73-88, 2016.

[127] K. Suzuki, F. Li, S. Sone and K. Doi, "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network," *IEEE Transactions on Medical Imaging,* vol. 24, no. 9, pp. 1138-1150, 2005.

[128] A. Jirapatnakul, A. Reeves, T. Apanasovich, A. Biancardi, D. Yankelevitz and C. Henschke, "Pulmonary nodule classification: size distribution issues," *ISBI,* pp. 1248-1251, April 2007.

[129] M. Aoyama, Q. Li, S. Katsuragawa, F. Li and S. Sone, "Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images," *Medical Physics,* vol. 30, no. 3, pp. 387-394, 2003.

[130] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation,* vol. 1, no. 4, pp. 541-551, 1989.

[131] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," *In Advances in neural information processing systems,* pp. 1097-1105, 2012.

[132] J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *CVPR,* pp. 248-255, June 2009.

[133] A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. van Riel, M. Wille, M. Naqibullah, C. Sánchez and B. van Ginneken, "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging,* vol. 35, no. 5, pp. 1160-1169, 2016.

[134] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint,* p. arXiv:1312.6229, 2013.

[135] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint ,* p. arXiv:1409.1556, 2014.

[136] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," *CVPR,* pp. 1-9, 2015.

[137] G. Litjens, T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken and C. Sánchez, "A survey on deep learning in medical image analysis," *arXiv preprint* , p. arXiv:1702.05747, 2017.

[138] H. Shin, H. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura and R. Summers, " Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging,* vol. 35, no. 5, p. 1285, 2016.

[139] B. van Ginneken, A. Setio, C. Jacobs and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," *ISBI,* April 2015.

[140] B. Huynh, H. Li and M. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging,* vol. 3, no. 3, p. 034501, 2016.

[141] N. Antropova, B. Huynh and M. Giger, "Performance comparison of deep learning and segmentation-based radiomic methods in the task of distinguishing benign and malignant breast lesions on DCE-MRI," *In SPIE Medical Imaging,* p. 101341G, March 2017.

[142] F. Ciompi, B. de Hoop, S. van Riel, K. Chung, E. Scholten, M. Oudkerk, P. de Jong, M. Prokop and B. van Ginneken, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Medical image analysis,* vol. 26, no. 1, pp. 195-202, 2015.

[143] M. Buty, Z. Xu, M. Gao, U. Bagci, A. Wu and D. Mollura, "Characterization of Lung Nodule Malignancy Using Hybrid Shape and Appearance Features," *In International Conference on Medical Image Computing and Computer-Assisted Intervention on Medical Image Computing and Computer-Assisted Intervention,* pp. 662-670, October 2016.

[144] H. Roth, L. Lu, A. Seff, K. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey and R. Summers, "A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations," *In International conference on medical image computing and computer-assisted intervention,* pp. 520-527, September 2014.

[145] W. Shen, M. Zhou, Y. F. C. Yang and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," *In International Conference on Information Processing in Medical Imaging,* pp. 588-599, June 2015.

[146] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang and J. Tian, "Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification," *Pattern Recognition,* vol. 61, pp. 663-673, 2017.

[147] S. Hamidian, B. Sahiner, N. Petrick and A. Pezeshk, "3D convolutional neural network for automatic detection of lung nodules in chest CT," *In SPIE Medical Imaging,* p. 1013409, March 2017.

[148] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. Mok, L. Shi and P. Heng, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE transactions on medical imaging,* vol. 35, no. 5, pp. 1182-1195, 2016.

[149] A. Mehrtash, A. Sedghi, M. Ghafoorian, M. Taghipour, C. Tempany, W. Wells, T. Kapur, P. Mousavi, P. Abolmaesumi and A. Fedorov, "Classification of Clinical Significance of MRI Prostate Findings Using 3D Convolutional Neural Networks," *In SPIE Medial Imaging*, p. 101342A, March 2017.

[150] J. Li, M. Fan, J. Zhang and L. Li, "Discriminating between Benign and Malignant Breast Tumors using 3D Convolutional Neural Network in Dynamic Contrast Enhanced-MR Images," *In SPIE Medical Imaging*, p. 1013808, March 2017.

[151] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424-432, October 2016.

[152] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus and A. Biller, "Deep MRI brain extraction: a 3D convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460-469, 2016.

[153] D. Kumar, A. Wong and D. Clausi, "Lung nodule classification using deep features in CT images," *CRV*, pp. 133-138, June 2015.

[154] R. Korez, B. Ibragimov, B. Likar, F. Pernuš and T. Vrtovec, "Intervertebral Disc Segmentation in MR Images with 3D Convolutional Networks," *In SPIE Medical Imaging*, p. 1013306, February 2017.

[155] R. Korez, B. Likar, F. Pernuš and T. Vrtovec, "Model-Based Segmentation of Vertebral Bodies from MR Images with 3D CNNs," *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 433-441, October 2016.

[156] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen and D. Comaniciu, "3D deep learning for efficient and robust landmark detection in volumetric data," *n International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 565-572, October 2015.

[157] A. Patel, S. van de Leemput, M. Prokop, B. van Ginneken and R. Manniesing, "Automatic cerebrospinal fluid segmentation in non-contrast CT images using a 3D convolutional network," *In SPIE Medical Imaging* , p. 1013420, March 2017.

[158] International Early Lung Cancer Action Program Investigators, "Survival of patients with stage I lung cancer detected on CT screening," *N Engl J Med*, vol. 2006, no. 355, pp. 1763-1771, 2006.

[159] National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *N Engl J Med*, vol. 2011, no. 365, pp. 395-409, 2011.

[160] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," *In Proceedings of the 23rd international conference on Machine learning* , pp. 233-240, June 2006.

[161] B. van Ginneken, S. Armato, B. de Hoop, S. van Amelsvoort-van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. R. Schilham, F. M. A. and N. Camarlinghi, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study," *Medical image analysis*, vol. 14, no. 6, pp. 707-722, 2010.

[162] I. Goodfellow, Y. Bengio, A. Courville and Y. Bengio, Deep learning, 1 ed., Cambridge: MIT press, 2016.

[163] J. Kittler, M. Hatef, R. Duin and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence,* vol. 20, no. 3, pp. 226-239, 1998.

[164] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning,* vol. 51, no. 2, pp. 181-207, 2003.

[165] S. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics,* no. 4, pp. 325-327, 1976.

[166] S. Le Cessie and J. Van Houwelingen, "Ridge estimators in logistic regression," *Applied statistics,* pp. 191-201, 1992.

[167] J. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters,* vol. 9, no. 3, pp. 293-300, 1999.

[168] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *In Aistats,* pp. 249-256, May 2010.

[169] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *In Proceedings of the 22nd ACM international conference on Multimedia,* pp. 675-678, November 2014.

[170] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[171] T. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence,* vol. 20, no. 8, pp. 832-844, 1998.

[172] Y. Freund, R. Iyer, R. Schapire and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of machine learning research,* vol. 4, no. Nov, pp. 933-969, 2003.

[173] G. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," *In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence,* pp. 338-345, August 1995.

[174] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg and J. Vanderplas, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, no. Oct, pp. 2825-2830, 2011.

[175] E. DeLong, D. DeLong and D. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics,* vol. 44, no. 3, pp. 837-845, 1988.

[176] A. Birnbaum, "Combining independent tests of significance," *Journal of the American Statistical Association,* vol. 49, no. 267, pp. 559-574, 1954.

[177] R. Fisher, Statistical methods for research workers, New York, NY: Springer, 1992.

[178] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez and M. Müller, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[179] S. Fotin, Y. Yin, H. Haldankar, J. Hoffmeister and S. Periaswamy, "Detection of soft tissue densities from digital breast tomosynthesis: Comparison of conventional and deep learning approaches," *In SPIE Medical Imaging ,* p. 97850X, March 2016.